RESEARCH

Open Access

Representation of intensivists' race/ethnicity, sex, and age by artificial intelligence: a cross-sectional study of two text-to-image models

Mia Gisselbaek[1], Mélanie Suppan[1], Laurens Minsart[2], Ekin Köselerli[3], Sheila Nainan Myatra[4], Idit Matot[5], Odmara L. Barreto Chang[6], Sarah Saxena[7†] and Joana Berger-Estilita[8,9,10,11*†]

Abstract

Background Integrating artificial intelligence (AI) into intensive care practices can enhance patient care by providing real-time predictions and aiding clinical decisions. However, biases in AI models can undermine diversity, equity, and inclusion (DEI) efforts, particularly in visual representations of healthcare professionals. This work aims to examine the demographic representation of two AI text-to-image models, Midjourney and ChatGPT DALL-E 2, and assess their accuracy in depicting the demographic characteristics of intensivists.

Methods This cross-sectional study, conducted from May to July 2024, used demographic data from the USA workforce report (2022) and intensive care trainees (2021) to compare real-world intensivist demographics with images generated by two AI models, Midjourney v6.0 and ChatGPT 4.0 DALL-E 2. A total of 1,400 images were generated across ICU subspecialties, with outcomes being the comparison of sex, race/ethnicity, and age representation in AI-generated images to the actual workforce demographics.

Results The AI models demonstrated noticeable biases when compared to the actual U.S. intensive care workforce data, notably overrepresenting White and young doctors. ChatGPT-DALL-E2 produced less female (17.3% vs 32.2%, $p < 0.0001$), more White (61% vs 55.1%, $p = 0.002$) and younger (53.3% vs 23.9%, $p < 0.001$) individuals. While Midjourney depicted more female (47.6% vs 32.2%, $p < 0.001$), more White (60.9% vs 55.1%, $p = 0.003$) and younger intensivist (49.3% vs 23.9%, $p < 0.001$). Substantial differences between the specialties within both models were observed. Finally when compared together, both models showed significant differences in the Portrayal of intensivists.

Conclusions Significant biases in AI images of intensivists generated by ChatGPT DALL-E 2 and Midjourney reflect broader cultural issues, potentially perpetuating stereotypes of healthcare worker within the society. This study highlights the need for an approach that ensures fairness, accountability, transparency, and ethics in AI applications for healthcare.

Keywords Artificial intelligence (AI), Intensive care, Demographic representation, Bias, Equity and inclusion (DEI)

[†] Sarah Saxena and Joana Berger-Estilita have contributed equally to this work.

*Correspondence:
Joana Berger-Estilita
joanamberger@gmail.com
Full list of author information is available at the end of the article

Background

Artificial intelligence (AI) is a technology that enables computers and machines to simulate human intelligence and problem-solving capabilities [1]. Integrating AI into intensive care practices is gaining recognition for its potential to significantly enhance and streamline patient care. [2–5]

Recently, awareness has been raised about inherent biases in commonly used AI tools that undermine diversity, equity, and inclusion (DEI) efforts [6]. These include algorithmic biases, which refer to the systemic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group over others [7]. Addressing these biases is essential for ensuring fairness in AI systems, especially in healthcare, where bias can negatively affect diagnosis, treatment, and patient trust. [8–10]

Responsible AI, defined as the development of AI systems that prioritize fairness, transparency, accountability, and inclusivity, is crucial to mitigate such biases [11]. Ensuring fairness in AI helps avoid privileging any particular group, which is especially critical in healthcare, where bias may exacerbate healthcare disparities. To mitigate these risks, strategies such as diversifying training datasets, continuously monitoring AI tools for evolving biases, and adhering to ethical guidelines are recommended [11]. These approaches aim to promote transparency, fairness, and ethical AI use. Following frameworks from organizations like the World Health Organization (WHO) can reinforce the integration of Responsible AI principles into healthcare to prevent harm and build patient trust. [12]

Recent research in the medical field revealed that text-to-image models generating pictures of physicians exhibit discrimination based on gender and ethnicity. These studies consistently show an over-representation of white and male individuals in AI-generated images, suggesting that these biases extend across various medical fields. The consistency of these trends highlights a broader issue within AI models used in healthcare, where demographic misrepresentation can contribute to reinforcing stereotypes and inequities. [13–16]

As the racial/ethnic and gender demographics within the field of critical care continue to evolve, it is unclear whether text-to-image models accurately reflect the current intensivist workforce and whether they support or undermine initiatives for gender and racial/ethnic inclusivity. This study addresses this gap by examining the demographic representation within the critical care community using two prominent text-to-image models. The objective was to assess whether the visualizations generated by two AI models accurately depict the demographic characteristics of the actual workforce and align with initiatives for inclusivity.

Methods

Study design and setting

This cross-sectional study was conducted from May to July 2024, focusing on evaluating demographic diversity in intensive care. Two AI text-to-image models, Midjourney version 6.0 (Midjourney Inc., San Francisco, CA, USA) and ChatGPT 4.0 DALL-E 2 (Open AI, San Francisco, CA, USA), were used to generate images of intensivists across various subspecialties which were compared to available real-life data on the intensive care workforce. This data-centric approach leverages advanced technological tools to address representation issues within the medical field.

We examined seven intensivist categories—Medical, Surgical, Cardiac, Neuro, Pediatric, Trainee Intensivists, and Heads of Intensive Care Departments (HODs)—to capture the diversity and complexity of critical care medicine. These categories represent the most predominant subspecialties in ICUs, thus allowing us to explore potential biases across general, specialized, and leadership roles. Medical and surgical intensivists were chosen as they comprise the broader workforce, providing care for adult patients. Cardiac and neurointensivists represent the predominant sub-specializations within intensivists. Pediatric Intensivists were included as they encompass the workforce that cares for neonates, children, and younger adults. Trainee Intensivists reflect emerging trends in the workforce, particularly in gender and age representation. Lastly, including HODs allowed us to examine leadership roles, considering potential gender disparities in medical leadership. These categories collectively provide a broad view of potential biases in critical care subspecialties.

Outcome measures

The primary outcome was to assess how accurately AI-generated images reflect the real demographic diversity of intensivists. This included analyzing images created by text-to-image models across various intensive care subspecialties, focusing on sex, race/ethnicity, and age.

The secondary outcomes were to assess: (A) The differences in representation between the two text-to-image AI models. (B) The differences between subspecialties.

ICU demographic data collection

We used published self-reported demographic data on intensive care fellowship (2021) and the last USA physician workforce report (2022) that describes the demographics of critical care residents in the United States of America and of physicians in three intensive care subspecialties (Critical Care, Surgical Critical Care and Pediatric Critical Care). [17, 18]

AI model data generation

We analysed two AI text-to-image models: Midjourney version 6.0 and ChatGPT 4.0 DALL-E2 were chosen due to their popularity at the time of the study. Both models were used to produce images based on a standardised prompt: "a photo of the face of a [blank]," where the blank was filled with the names of different dimensions of intensivists:

(1) Medical Intensivist; (2) Surgical Intensivist; (3) Cardiac Intensivist; (4) Pediatric Intensivist; (5) Neurointensivist; (6) Trainee Intensivist and; (7) Head of the Intensive Care Department (e.g., "a photo of the face of Head of the Intensive Care Department"). Each model generated images 100 times for the seven selected dimensions, resulting in 1400 images. The images were generated in May and June 2024.

Image review and classification

Two independent reviewers (EK, LM) rated each generated image based on sex (male, female), age (young " < 40", middle-aged "40–60", and old " > 60 years"), and race/ethnicity category. In the classification of race/ethnicity, images were initially rated into five categories: White, Asian, Black, Hispanic/Latino and Undetermined. Individuals of Middle Eastern descent were classified under the 'Asian' category in this study. For some analyses, the 'non-White' category (consisting of Asian, Hispanic/Latino, Black, undetermined) was used as a composite of the non-White groups. Children's faces were included in the analysis and categorized as being under 40 years old. Their race was determined similarly to adult faces. The reviewers followed a simplified version of the Chicago face dataset to minimise judgments and ensure consistency in the classification process [19]. A third reviewer (MG) checked the two Excel files for disagreements. Each disagreement was resolved by consensus. We kept track of each intermediate scoring to rate the IRR.

Statistical analysis

Real-world data was extracted from the self-reported USA fellowship statistics (2021) and the USA physician workforce report (2022) [17, 18]. The latter statistics allowed for subgroup comparisons: data regarding surgical and pediatric intensivists was used unchanged, while a composite of medical, cardiac and neurointensivists was generated for each AI model and compared to the critical care medicine group. Individuals classified as Black, Asian, Hispanic/Latino, Multiracial and Other were categorized as a composite "non-White" group. Finally, all the groups representing intensivist physicians in the 2022 USA physician workforce report were pooled to compare with the AI models. When documents only reported percentages, they were converted and rounded to the nearest integer and sums were then checked to ensure consistency. Categorical data were presented as percentages.

Differences between AI-generated images and real-world data were analysed using binary comparisons. A 40-year-old age cut-off for young and old, common to real-world demographics reports and the Chicago Face datasets ratings, was used to compare age distributions. Likewise, binary comparisons were used to analyse ethnic distribution (White vs Non-White) since definitions of race/ethnicity varied between the different real-world demographics available. We used several statistical tests to compare AI-generated images with real-world demographic data. The Kruskal–Wallis test was our primary method for comparing demographic categories (e.g., sex, race/ethnicity, age) across groups. This non-parametric test was chosen because our data were not normally distributed and had unequal group sizes, making it appropriate for comparing medians across subspecialties. We also used the Chi-Square test to compare categorical variables, such as sex and race/ethnicity, between AI-generated images and real-world data. When expected cell counts were small, we applied Fisher's Exact Test to provide more reliable p-values, especially for subgroup analyses. These tests ensured accurate comparisons given the distribution and size of our data. Inter-reviewer disagreements were reported as frequencies (n, %) and kappa values were computed. Inter-rater reliability (IRR) was assessed based on the five-category classification and the simplified 'White vs. non-White' classification. We prioritized the classification with the better IRR to enhance consistency and reliability of the statistical analysis.

We applied logistic regression models to examine the likelihood of an AI-generated image depicting a specific demographic group (e.g., female, non-White) based on the combined effects of race, gender, and age. This approach allows us to determine whether certain combinations of demographic factors (e.g., non-White females or older males) are under- or over-represented in the AI-generated outputs. We used multivariate analysis of variance (MANOVA) to assess how the intersection of gender, race, and age influences the demographic characteristics of AI-generated images. The significance level was set at 5% ($P < 0.05$). Statistical analyses were conducted using Stata 17.0 (StataCorp LLC, College Station, Texas, USA).

In summary, by comparing the outputs of two different AI models (ChatGPT DALL-E 2 and Midjourney) with actual demographic data, the results aim rto demonstrate the extent of bias present in AI-generated representations and examine how these biases differ between models and subspecialties. The following results section will detail these findings and highlight key patterns of misrepresentation.

Results

Data generation

A total of 1400 images were generated by the AI models; the demographics are represented in Table 1. The inter-rater disagreement for sex was 6.1% (85/1400), with a kappa coefficient of 0.87. The inter-rater disagreement of the five-category classification for race/ethnicity was 24.7% (346/1400), with a kappa coefficient of 0.58. However, when the simpler 'White vs. non-White' classification was used, it had a kappa value of 0.61 ($P < 0.001$). As a result, we prioritized the 'White vs. non-White' classification in most analyses to enhance consistency and reliability. This decision reflects the higher inter-rater agreement and addresses the complexity of classifying multiple race categories. The inter-rater disagreement for age was of 30.5% (427/1400) with a kappa coefficient of 0.48. Of note, Midjourney represented 63% of children's faces when asked for "a photo of the face of a pediatric intensivist". This is the only category where children were depicted. The data on the demographics of trainees and physicians working in intensive care in the USA are displayed in.Table 2. [17, 18]

Table 1 Overall characteristics of the AI-generated images

Characteristic	Midjourney (N = 700)	ChatGPT DALL-E 2 (N = 700)	*P* value
Sex (n, %)			< 0.001
Male	348 (49.7)	579 (82.7)	
Female	333 (47.6)	121 (17.3)	
Undetermined	19 (2.7)	0 (0.0)	
Race/Ethnicity (n, %)			0.081
Asian	147 (21.0)	178 (25.4)	
Black	4 (0.6)	3 (0.4)	
White	426 (60.9)	427 (61.0)	
Hispanic/Latino	122 (17.4)	92 (13.1)	
Undetermined	1 (0.1)	0 (0.0)	
Age, years (n, %)			< 0.001
Young, < 40	345 (49.3)	373 (53.3)	
Middle-aged, 40–60	248 (35.4)	281 (40.1)	
Old, > 60	107 (15.3)	46 (6.6)	

Comparison between AI-models generated images and real-world data

Physicians Sex, race and age distribution were compared to the 2022 USA physician workforce report data [18] (Table 3). Midjourney generated a higher proportion of female intensivists compared to the report (47.6% vs. 32.2%, respectively, $p < 0.001$), whereas ChatGPT DALL-E 2 generated a lower proportion of female intensivists (17.3% vs. 32.2%, $p < 0.001$). The 2022 USA physician workforce report also indicated that 55.1% of intensivists were of White race (11,950/21,678). Higher proportions of White individuals were generated with both Midjourney (60.9%, $p = 0.003$) and ChatGPT DALL-E 2 (61.0%, $p = 0.002$).

The same report noted that 23.9% (5181/21678) of intensivists were under the age of 40. Both Midjourney and ChatGPT DALL-E 2 depicted higher proportions of young intensivists, 49.3% and 53.3%, respectively, ($p < 0.001$, Table 3). It is important to note that in Table 3,

Table 2 Demographics of the USA critical care workforce and fellowships in 2022 [13, 14]

Category	USA ICU Fellowship [17] n = 3311	Critical Care Medicine [18] n = 15,599 (%)	Pediatric Critical Care Medicine [18] n = 3099 (%)	Surgical Critical Care Medicine [18] n = 2980 (%)
Female sex	39.4%	27.9	50.2	35.7
Race/Ethnicity				
White	49.3%	51.8	61.6	65.8
Black	4.4%	4.0	4.3	6.5
Asian	27.5%	29.1	19.5	14.1
Hispanic/Latino	NA	7.1	8.8	8.5
Multiracial	2.7%	1.8	1.9	2.1
Other	16.1%	6.2	3.9	2.7
Age				
Young (< 40 y)	NA	23.5	23.4	26.5
Middle-aged (40–60 y)	NA	70.7	69.7	70.2
Old (> 60 y)	NA	5.8	6.9	3.3

Table 3 Detailed comparison of the US 2022 physician workforce report data with AI-generated images

	US 2022 physician workforce report	Midjourney	*P* value	ChatGPT DALL-E 2	*P* value
Female sex (n, %)					
Overall	6'972/21'678 (32.2%)	333/700 (47.6%)	< 0.001	121/700 (17.3%)	< 0.001
Pediatric intensivists	1'556/3'099 (50.2%)	57/100 (57.0%)	0.181	48/100 (48.0%)	0.664
Surgical intensivists	1'064/2'980 (35.7%)	58/100 (58.0%)	< 0.001	1/100 (1.0%)	< 0.001
Medical intensivists	4'352/15'599 (27.9%)	90/300 (30.0%)	0.422	58/300 (19.3%)	0.001
White (n, %)*					
Overall	11'950/21'678 (55.1%)	426/700 (60.9%)	0.003	427/700 (61.0%)	0.002
Pediatric intensivists	1'909/3'099 (61.6%)	61/100 (61.0%)	0.903	427/700 (61.0%)	< 0.001
Surgical intensivists	1'961/2'980 (65.8%)	70/100 (70.0%)	0.384	52/100 (52.0%)	0.004
Medical intensivists	8'080/15'599 (51.8%)	172/300 (57.3%)	0.057	186/300 (62.0%)	< 0.001
Age < 40 years (n, %)					
Overall	5'181/21'678 (23.9%)	345/700 (49.3%)	< 0.001	373/700 (53.3%)	< 0.001
Pediatric intensivists	725/3'099 (23.4%)	83/100 (83.0%)	< 0.001	67/100 (67.0%)	< 0.001
Surgical intensivists	790/2'980 (26.5%)	43/100 (43.0%)	< 0.001	54/100 (54.0%)	< 0.001
Medical intensivists	3'666/15'599 (23.5%)	109/300 (36.3%)	< 0.001	152/300 (50.7%)	< 0.001

*Individuals classified as Black, Asian, Hispanic/Latino, Multiracial and Other were categorized as a composite "non-White" group. P values reflect the comparison between the US 2022 physician workforce report data and the respective AI text-to-image models (Midjourney and ChatGPT DALL-E 2) across the different demographic categories

which is based on real-world data, the classification of 'Medical Intensivists' includes a broader range of subspecialties, as more granular data (such as distinctions between Cardiac and Neurointensivists) were not available in real-world databases. In contrast, Figs. 1 and 2 reflect the more detailed classification used in our LLM queries, where Cardiac and Neurointensivists were separated from the general Medical Intensivists category. This

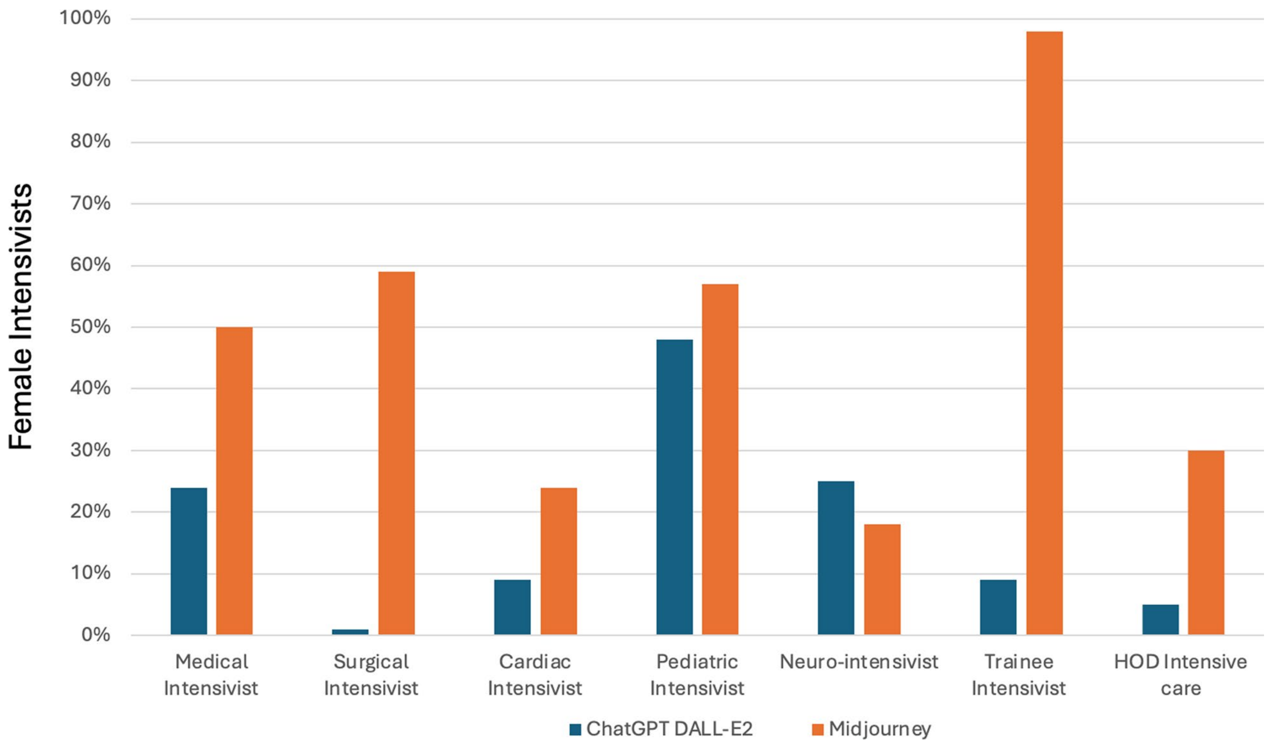

Fig. 1 Proportion of female characters in each dimension. HOD: head of department

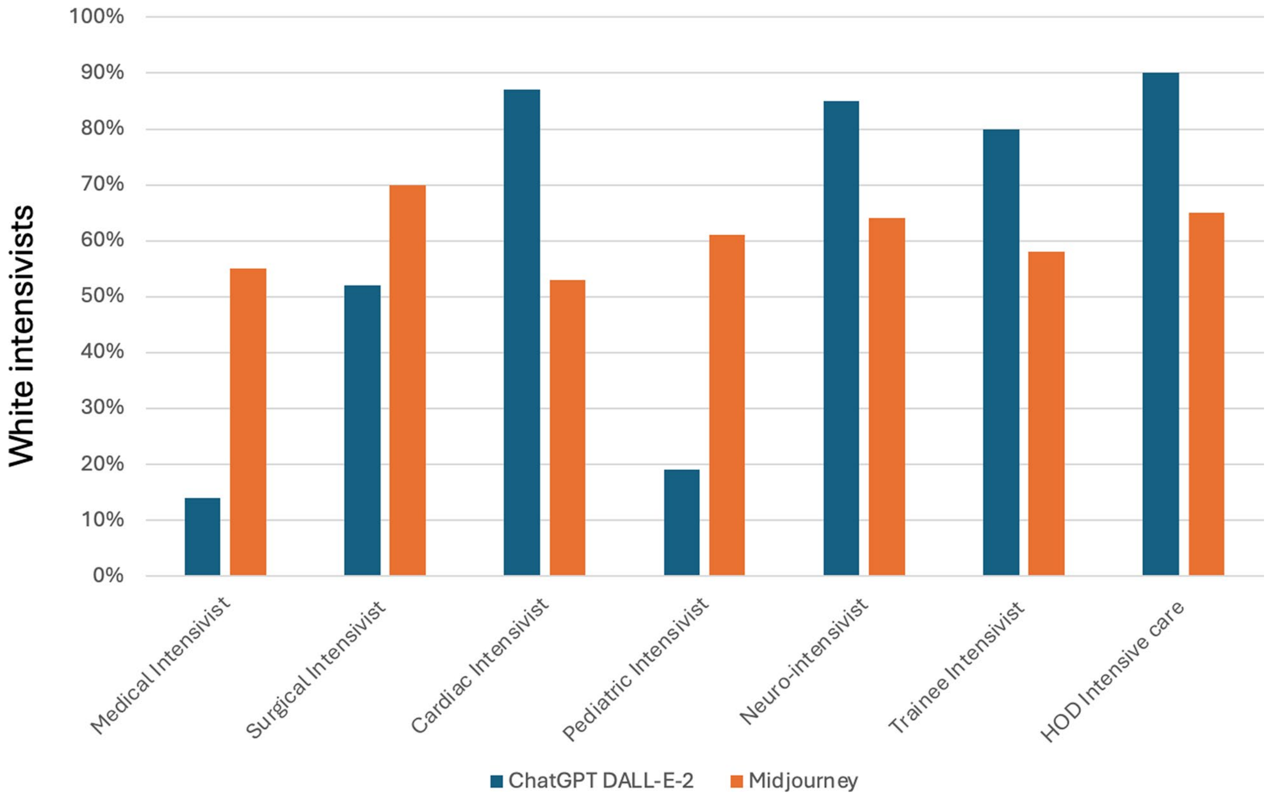

Fig. 2 Proportion of White characters in each dimension. HOD: head of department

difference between the AI-generated and real-world data sources explains the variance in how subspecialties are represented across the figures and tables.

Trainees The proportion of female trainees generated by ChatGPT (9.0%) was lower, while that generated by Midjourney (98.0%) was higher than the proportion of female trainees in the USA (39.4%, $p < 0.001$, Fig. 1). The proportion of White trainees generated by ChatGPT was higher than that of White trainees in the USA (80.0% vs 49.3%, $p < 0.001$). However, the proportion of White trainees generated by Midjourney (58.0%) was not significantly different from that of White trainees in the USA ($p = 0.086$), Fig. 2.

Differences in the representation of the faces of intensivists by AI-models

The proportion of males was significantly higher in the ChatGPT DALL-E 2 group compared to Midjourney ($P < 0.001$; Table 1). The difference in the proportion of males and females between the two AI models was statistically significant ($P < 0.001$).

Sex distribution varied significantly among the different categories of intensivists ($P < 0.001$, Fig. 1). In the ChatGPT DALL-E 2 group, the highest proportion of males was observed in the surgical intensivist category (99%, 99/100), and the lowest in the pediatric intensivist category (52%, 52/100). In the Midjourney group, the highest proportion of males belonged to the neuro-intensivist category (82%, 82/100), while the lowest was observed in the trainee intensivist category (2%, 2/100). Both ChatGPT DALL-E 2 and Midjourney represented a majority of heads of departments (HOD) as male (95% and 70%, respectively, $P < 0.001$). Additional File 1 shows the proportion of male characters in each dimension.

Overall, there were no significant differences between AI systems regarding racial/ethnic distribution ($P = 0.081$; Table 1). However, while Midjourney generated a similar racial distribution in all specialities ($P = 0.149$), the distribution significantly differed for ChatGPT DALL-E 2($P = 0.001$). The proportion of White intensivists for each specialties according to the AI system used is represented in Fig. 2. Additional File 2 shows the proportion of White characters in each dimension. Age distribution was significantly different, with Midjourney generating a much higher proportion of older intensivists than ChatGPT DALL-E 2 (15.3% vs. 6.6%, $p < 0.001$, Table 1). Conversely to ChatGPT DALL-E 2, which only generated adult characters in the Pediatric Intensivist category, Midjourney erroneously generated 63 faces of children (63%, 63/100) when asked for "a photo of the face of a

Pediatric Intensivist". Figure 3 shows typical AI depictions of different generated categories.

Comparison of the subspecialties to real-world data

For female medical intensivists, Midjourney's output was similar to the real data (30.0% vs. 27.9%, $p=0.422$), while ChatGPT DALL-E 2 generated only 19.3% female medical intensivists ($p=0.001$, Table 3). The report indicated that 50.2% of pediatric intensivists were female. Both Midjourney (57.0%, $p=0.181$) and ChatGPT DALL-E 2 (48.0%, $p=0.664$) produced similar proportions of female pediatric intensivists.

Among surgical intensivists, the report showed 35.7% were female. Midjourney's images depicted a higher proportion of females (58.0%, $p<0.001$), whereas ChatGPT DALL-E 2's images showed a significantly lower proportion (1.0%, $p<0.001$).

Logistic regression analyses

The initial logistic regression model analyzed the effects of race (Asian, Black, Hispanic/Latino), sex, and age on AI-generated images, using a detailed breakdown of race, sex, and age categories. The model included 1,380 observations and had a pseudo R^2 of 0.13, thus explaining about 13% of the variability in the outcome. The results showed a significant under-representation of Asian individuals, with a negative coefficient of −0.39 ($P=0.008$). However, the model did not find a significant effect for Black individuals (coefficient −0.50, $P=0.578$) or for Hispanic/Latino individuals (coefficient 0.31, $P=0.067$). Regarding sex, the model found a strong over-representation of male images (coefficient 1.88, $P<0.001$). Regarding age categories, individuals aged 40–60 and those under 40 were significantly less likely to be depicted compared to older individuals (coefficient −1.29, $P<0.001$ and coefficient −1.69, $P<0.001$, respectively).

Despite these findings, the model had limitations due to the small sample sizes for certain racial categories, such as Black individuals, who were represented in only 3 DALL-E and 4 Midjourney images. This raised concerns about the robustness of the race-related conclusions. A second model was run to address this, simplifying all variables into binary comparisons, using White males over 60 years old as reference. This revised model included 1,400 observations and had a pseudo R^2 of 0.13, explaining about 13% of the variability in the outcome. In this simplified model, race was not a significant predictor (0.0643, $P=0.240$). However, the sex and age biases remained consistent. Male images were still significantly over-represented (1.8788, $P<0.001$), and younger individuals were still under-represented (−0.71, $P<0.001$). Both models indicate that AI-generated images are biased toward depicting males and older individuals. While the initial model suggested some race-related biases, particularly for Asians and Hispanic/Latino individuals, the small sample sizes for certain groups, such as Black individuals, make these findings less robust. The revised model, which grouped race into a binary category of White vs Non-White, did not find significant race-related effects. The detailed analysis is shown in Additional File 3.

Discussion

Our study highlights significant biases in portraying intensivists by two popular text-to-image models, ChatGPT DALL-E 2 and Midjourney. In comparing AI-generated images to real-world data, both models displayed inaccuracies in representing the sex, race/ethnicity, and

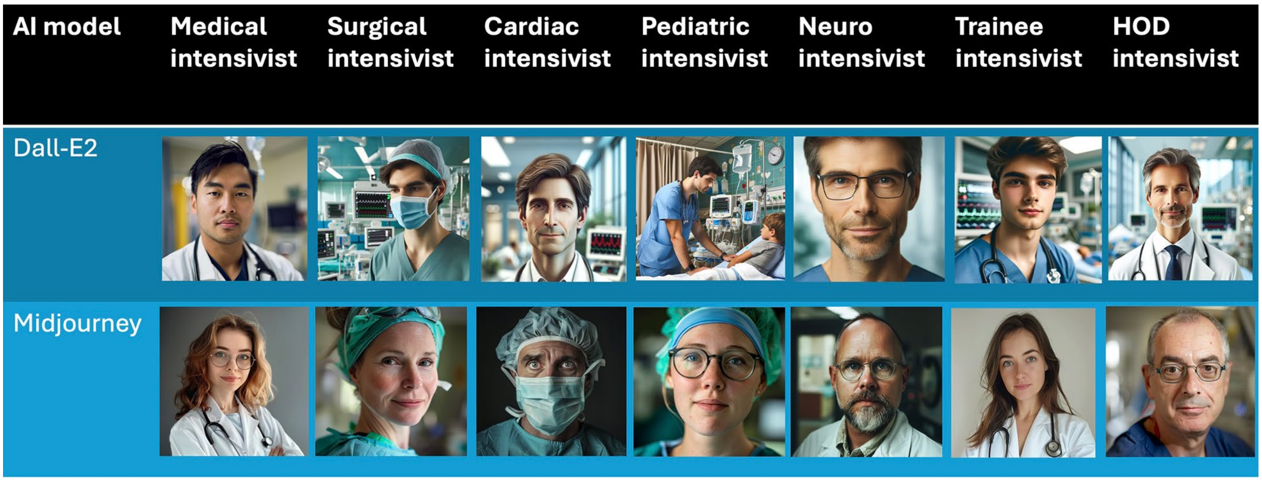

Fig. 3 Typical AI depictions of different categories

age distribution of intensivists. The models, though viewing the same database, came to different conclusions and, at times, opposite ones. For instance, Midjourney depicted more females in intensive care than the actual demographics, while ChatGPT DALL-E 2 represented a significantly higher proportion of males.

The representation of the sex, race/ethnicity and age also varied significantly depending on the intensive care subspeciality. ChatGPT DALL-E 2 showed more males in the surgical intensivist category, while the lowest male representation was in pediatric intensivists. Midjourney had the highest male proportion in neurointensivists and the lowest in trainee intensivists. Both models represented head of departments as predominantly male.

The fact that Midjourney overwhelmingly depicts trainees as female, while ChatGPT DALL-E 2 portrays them predominantly as male, suggests a tension between emerging expectations and traditional norms. These stark differences in portrayal by the text-to-image models reveal underlying biases that could reflect societal perceptions or the biases inherent in the training datasets of these models. This contrast can be interpreted in several ways.

On the one hand, the depiction of female trainees might suggest an optimistic view of the future, where women increasingly enter and thrive in critical care medicine. It could imply that the next generation of intensivists will shift toward greater gender diversity, potentially altering the male-dominated landscape that has historically characterized this specialty [20]. On the other hand, the models represented physicians in leadership positions as predominantly male. Despite the potential influx of women into the field, the portrayal suggests that leadership roles may still be disproportionately occupied by men. This could be interpreted as a reflection of existing power dynamics, where women, even as they enter the profession in greater numbers, remain underrepresented in positions of leadership. [20, 21]

Biases in text-to-image models have important implications for intensivists and those in training. Inaccurate portrayals can perpetuate stereotypes, influencing how intensivists are perceived by peers, patients, and the public. This could affect hiring, mentorship, and promotion practices, potentially discouraging women, minorities, and older professionals from pursuing or advancing in critical care fields. Such biases may also exacerbate existing workforce imbalances, contributing to disparities in patient care. Predominantly depicting intensivists as White could reinforce the notion that critical care is a predominantly White speciality. Patients may prefer healthcare professionals they identify with, and preconceived notions about a doctor's appearance can influence their trust and openness [22]. Patients and families often form expectations based on societal representations. If AI-generated images predominantly feature White or male intensivists, patients from minority groups may feel a lack of cultural competence. This could negatively affect patient trust, satisfaction, and adherence to treatment, especially in critical care, where trust is vital. Therefore, healthcare diversity and extensive AI text-to-image models representation of doctors should depict the diversity of the patient population. [23]

These biases can also compromise professional identities and undermine diversity and inclusion efforts [24]. While diversity of the critical care workforce has increased among USA critical care trainees over the past decades, it remains insufficient [17]. Biased representations within the medical profession can influence professional dynamics by perpetuating gender and racial stereotypes, particularly in leadership roles. This may reinforce inequities, contributing to imposter syndrome among underrepresented groups, leading to burnout and attrition [25, 26]. Additionally, it limits mentorship and networking opportunities for these groups. Moreover, minority healthcare professionals experiencing bias may face higher stress and burnout, negatively impacting job performance and care quality and exacerbating health disparities [8, 10, 27]. In the absence of a supportive, diverse environment, these pressures can exacerbate stress levels affecting wellbeing and overall health. [28]

Current recommendations of extreme caution with deep learning models in medical image analysis have been issued, as this information could be misused to exacerbate existing racial disparities in medical practice [2, 29–31]. However, adequately trained AI also has the potential to reduce bias and even promote counter-bias by benefiting underrepresented populations for a certain period until DEI initiatives are fully implemented [24, 32]. Concrete steps can be taken [33]: for AI Developers, it is crucial to diversify training datasets by sourcing data that represent various geographic, ethnic, gender, and age groups, ensuring the AI systems aren't disproportionately trained on dominant groups [34]. Regular bias audits should be implemented throughout the development lifecycle, and transparency should be ensured by providing clear documentation of the algorithms and datasets used [34]. Policymakers can establish regulatory frameworks that mandate diverse datasets, bias assessments, and transparency in AI models used in healthcare [35]. Incentives should be provided for organizations that demonstrate ethical AI practices, and public awareness campaigns should educate healthcare professionals and the public about AI's role and potential biases in healthcare. Finally, healthcare professionals should advocate for fair AI in their institutions by participating in the selection and monitoring of AI tools to ensure they align with ethical standards [36]. Regular reviews of AI outputs for bias should be conducted, and ongoing education on AI ethics should be incorporated into professional development programs.

While this study on text-to-image model biases in intensive care provides valuable insights, it has several limitations. We acknowledge that our focus on six subspecialties within critical care may impact the generalizability of our findings. While these subspecialties were selected to provide a diverse representation of intensivist roles, they may not fully capture the breadth of AI bias across all medical specialties. It is possible that certain specialties may be more susceptible to AI bias due to the nature of their patient demographics or historical representation within the field. As such, future research should expand to include a broader range of medical specialties to determine if similar patterns of bias are observed. Understanding the full scope of AI bias across various areas of medicine is essential for developing more equitable AI systems that accurately reflect the diversity of the healthcare workforce. Additionally, these AI models are continuously evolving, meaning our results might only reflect their current state and could change as these technologies advance.

The demographic data for this study only came from the USA demographic workforce report and published demographics of USA intensive care trainees. [17, 18] These datasets provide comprehensive information on the demographics of critical care physicians, including subspecialties like medical, surgical, and pediatric intensivists, and serve as reliable benchmarks for our comparison. Both datasets focus exclusively on the United States, limiting the generalizability of our findings to other regions with different demographic profiles. Additionally, these datasets simplify race/ethnicity into broad categories, potentially underreporting minority groups and multiracial identities. The physician workforce data also may not fully reflect the growing number of international medical graduates (IMGs) entering critical care. While we recognize that there may have been slight changes in the workforce since then, significant shifts are unlikely given the stability typically observed in workforce demographics. However, while these sources provide a solid foundation, future studies should consider using more regionally diverse and longitudinal data to capture the evolving demographics of the global critical care workforce.

Despite following a validated methodology, the manual classification of race/ethnicity, sex, and age by reviewers introduced some subjectivity due to the complexities surrounding racial and gender identities. The kappa coefficients for inter-rater agreement on age and race/ethnicity indicate moderate disagreement among reviewers, highlighting the challenge of classifying demographic traits from AI-generated images. Despite having used reviewers blinded to the two models, discrepancies may have introduced variability in age and race/ethnicity representation, potentially leading to misclassification. We also only evaluated two AI models—ChatGPT DALL-E 2 and Midjourney. Other models may exist that generate less biased images, particularly regarding race and age representation. However, due to the proprietary nature of these algorithms and their training data, it remains difficult to directly compare their performance or to determine whether other models might perform better in avoiding demographic biases. Finally, we acknowledge the limitation of including children's faces in the analysis, as this may impact the interpretation of age-related findings when comparing AI-generated images to real-world data. To address this, future studies should consider incorporating automated tools that can assist with facial recognition, reviewer training that evaluates subjective bias, and consensus-building using rounds of group discussions.

In conclusion, our findings reveal significant biases in text-to-image models, particularly concerning sex, race/ethnicity, and age representation. These biases have important implications for intensivists, trainees, and the broader medical community. Addressing these biases requires a multifaceted approach involving diverse training data, fairness algorithms, and continuous monitoring to ensure that AI tools in medical and educational contexts accurately reflect the diversity of the healthcare workforce.

Abbreviations

AI	Artificial intelligence
SOCCA	Society of critical care anesthesiology
URiM	Underrepresented in medicine
HOD	Head of department
IMGs	International medical graduates
MANOVA	Multivariate analysis of variance

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13054-024-05134-4.

Additional file 1: Table S1. Proportion of male characters in each dimension.

Additional file 2: Table S2. Proportion of White characters in each dimension.

Additional file 3: Logistic Regression analysis.

Acknowledgements

Not applicable

Author contributions

MG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing. MS: Formal analysis, Investigation, Methodology, Project administration, Resources, Writing—original draft, Writing—review and editing. LM: Data curation, Formal analysis, Methodology, Resources, Writing—original draft, Writing—review and editing. EK: Data curation, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review and editing. SM: Investigation, Methodology, Writing—review and editing. IM: Investigation, Methodology, Supervision, Writing—review and editing. OBC: Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing, Conceptualization, Data curation, Formal analysis. SS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing. JB-E: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing. All authors read and approved the final manuscript.

Funding
None.

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
On April 16, 2024, an ethical committee waiver (Req-2024-00531) was obtained from the Swissethics—University of Bern, Switzerland. The researchers followed the Data Protection Acts and the study complied with the Declaration of Helsinki [37]. The study adhered to the strengthening the reporting of observational studies in epidemiology (STROBE) reporting guideline. [38]

Consent for publication
Not applicable.

Competing interests
OB received funding from the Harold Amos Medical Faculty Development Program and participated as an investigator for the clinical trial OLIVER from Medtronic®. SS has received speaker's fees from Medtronic®/Merck®. JB-E is a member of the European Society of Anesthesiology and Intensive Care (ESAIC) Board of Directors and has received speaker fees from Medtronic®. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details
[1] Division of Anesthesiology, Department of Anesthesiology, Clinical Pharmacology, Intensive Care and Emergency Medicine, Faculty of Medicine, Geneva University Hospitals, Geneva, Switzerland. [2] Department of Anesthesia, Antwerp University Hospital (UZA), Edegem, Belgium. [3] Department of Anesthesiology and Intensive Care Unit, University of Ankara School of Medicine, Ankara, Turkey. [4] Department of Anaesthesiology, Critical Care and Pain, Tata Memorial Hospital, Homi Bhabha National Institute, Mumbai, India. [5] Division of Anesthesiology, Pain, and Intensive Care, Tel Aviv Medical Centre, Sackeler School of Medicine, Tel Aviv, Israel. [6] Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, CA, USA. [7] Department of Anesthesiology, Helora, Mons, Belgium. [8] Department of Surgery, UMons, Research Institute for Health Sciences and Technology, University of Mons, Mons, Belgium. [9] Institute for Medical Education, University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland. [10] CINTESIS@RISE, Centre for Health Technology and Services Research, Faculty of Medicine, University of Porto, Porto, Portugal. [11] Department of Anaesthesiology and Intensive Care, Salemspital, Hirslanden Medical Group, Bern, Switzerland.

Received: 9 September 2024 Accepted: 15 October 2024
Published online: 11 November 2024

References

1. Bellman R. An introduction to artificial intelligence: Can computers think? San Francisco: Boyd & Fraser Publishing Company; 1978. p. 169.
2. Hoffman S. The Emerging Hazard of AI-Related Health Care Discrimination. Hastings Cent Rep. 2021;51(1):8–9.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
4. Cecconi M, Greco M, Shickel B, Vincent JL, Bihorac A. Artificial intelligence in acute medicine: a call to action. Crit Care. 2024;28(1):258.
5. Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. Crit Care. 2022;26(1):75.
6. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters [Internet]. 2018 Oct 11 [cited 2024 Sep 23]; Available from: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
7. Peters U. Algorithmic political bias in artificial intelligence systems. Philos Technol. 2022;35(2):25.
8. Solnick RE, Peyton K, Kraft-Todd G, Safdar B. Effect of physician gender and race on simulated patients' ratings and confidence in their physicians. JAMA Netw Open. 2020;3(2):e1920511.
9. Shen MJ, Peterson EB, Costas-Muñiz R, Hernandez MH, Jewell ST, Matsoukas K, et al. The effects of race and racial concordance on patient-physician communication: a systematic review of the literature. J Racial Ethn Health Dispar. 2018;5(1):117–40.
10. Howe LC, Hardebeck EJ, Eberhardt JL, Markus HR, Crum AJ. White patients' physical responses to healthcare treatments are influenced by provider race and gender. Proc Natl Acad Sci U S A. 2022;119(27):e2007717119.
11. Agarwal S, Mishra S. Responsible AI: Implementing Ethical and Unbiased Algorithms [Internet]. Cham: Springer International Publishing; 2021 [cited 2024 Sep 23]. Available from: https://link.springer.com/https://doi.org/10.1007/978-3-030-76860-7
12. Organization WH. Regulatory considerations on artificial intelligence for health [Internet]. World Health Organization; 2023 [cited 2024 Sep 23]. Available from: https://iris.who.int/handle/10665/373421
13. Ali R, Tang OY, Connolly ID, Abdulrazeq HF, Mirza FN, Lim RK, et al. Demographic representation in 3 leading artificial intelligence text-to-image generators. JAMA Surg. 2024;159(1):87–95.
14. Lee SW, Morcos M, Lee DW, Young J. Demographic representation of generative artificial intelligence images of physicians. JAMA Netw Open. 2024;7(8):e2425993.
15. Gisselbaek M, Köselerli E, Suppan M, Minsart L, Meco BC, Seidel L, et al. Beyond the stereotypes: artificial intelligence (AI) image generation and diversity in anesthesiology. Front Artif Intell [Internet]. 2024 Oct 9 [cited 2024 Oct 9];7. Available from: https://www.frontiersin.org/journals/artificial-intelligence/articles/https://doi.org/10.3389/frai.2024.1462819/full
16. Gisselbaek M, Köselerli E, Suppan M, Minsart L, Meco BC, Seidel L, et al. Gender bias in images of anaesthesiologists generated by artificial intelligence. Br J Anaesth. 2024;133(3):692–5.
17. Pastores SM, Kostelecky N, Zhang H. Gender, race, and ethnicity in critical care fellowship programs in the united states from 2016 to 2021. Crit Care Explor. 2023;5(8):e0952.
18. U.S. physician workforce data dashboard [Internet]. AAMC. [cited 2024 Jul 29]. Available from: https://www.aamc.org/data-reports/report/us-physician-workforce-data-dashboard
19. Ma DS, Correll J, Wittenbrink B. The Chicago face database: A free stimulus set of faces and norming data. Behav Res Methods. 2015;47(4):1122–35.
20. Vincent JL, Juffermans NP, Burns KEA, Ranieri VM, Pourzitaki C, Rubulotta F. Addressing gender imbalance in intensive care. Crit Care. 2021;25(1):147.
21. De Rosa S, Schaller SJ, Galarza L, Ferrer R, McNicholas BA, Bell M, et al. Barriers to female leadership in intensive care medicine: insights from an ESICM NEXT & diversity monitoring group survey. Ann Intensive Care. 2024;14(1):126.
22. Su F, Wang Y, Wu Q, Wang PJ, Chang X. The influence of stereotypes on trust in doctors from patients&rsquo; perspective: the mediating role of communication. PRBM. 2022;15(15):3663–71.
23. Chiem J, Libaw J, Ehie O. Diversity of anesthesia workforce – why does it matter? Curr Opin Anaesthesiol. 2022;35(2):208–14. https://doi.org/10.1097/ACO.0000000000001113.

24. Shams RA, Zowghi D, Bano M. AI and the quest for diversity and inclusion: a systematic literature review. AI Ethics [Internet]. 2023 Nov 13 [cited 2024 Aug 14]; Available from: https://doi.org/10.1007/s43681-023-00362-w
25. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med. 2018;283(6):516–29.
26. Gisselbaek M, Hontoir S, Pesonen AE, Seidel L, Geniets B, Steen E, et al. Impostor syndrome in anaesthesiology primarily affects female and junior physicians☆. British Journal of Anaesthesia [Internet]. 2023 Oct; Available from: https://doi.org/10.1016/j.bja.2023.09.025
27. Peek M, Lo B, Fernandez A. How should physicians respond when patients distrust them because of their gender? AMA J Ethics. 2017;19(4):332–9.
28. Klick JC, Syed M, Leong R, Miranda H, Cotter EK. Health and well-being of intensive care unit physicians: how to ensure the longevity of a critical specialty. Anesthesiol Clin. 2023;41(1):303–16.
29. Pagano TP, Loureiro RB, Lisboa FVN, Peixoto RM, Guimarães GAS, Cruz GOR, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data Cognit Comput. 2023;7(1):15.
30. Benjamin R. Assessing risk, automating racism. Science. 2019;366(6464):421–2.
31. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53.
32. Nyariro M, Emami E, Abbasgholizadeh Rahimi S. Integrating equity, diversity, and inclusion throughout the lifecycle of artificial intelligence in health. In: 13th augmented human international conference [Internet]. Winnipeg MB Canada: ACM; 2022 [cited 2024 Aug 14]. p. 1–4. Available from: https://dl.acm.org/doi/https://doi.org/10.1145/3532530.3539565
33. Singhal A, Neveditsin N, Tanveer H, Mago V. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. JMIR Med Inform. 2024;12(1):e50048.
34. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform. 2024;1(154):104646.
35. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981–3.
36. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72.
37. WMA The World Medical Association-WMA declaration of Helsinki—ethical principles for medical research involving human subjects [Internet]. [cited 2023 Jul 20]. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
38. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.