Original Paper

Artificial Intelligence Bias in Health Care: Web-Based Survey

Carina Nina Vorisek[1], MSc, MD; Caroline Stellmach[1], MSc; Paula Josephine Mayer[1]; Sophie Anne Ines Klopfenstein[1,2], MD; Dominik Martin Bures[3], MA; Anke Diehl[3], MA, MD; Maike Henningsen[4], Prof Dr; Kerstin Ritter[5], Prof Dr; Sylvia Thun[1], Prof Dr

[1]Core Facility Digital Medicine and Interoperability, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany

[2]Institute for Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany

[3]Stabsstelle Digitale Transformation, Universitätsmedizin Essen, Essen, Germany

[4]Faculty of Health, University of Witten/Herdecke, Witten, Germany

[5]Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin, Germany

Corresponding Author:
Carina Nina Vorisek, MSc, MD
Core Facility Digital Medicine and Interoperability
Berlin Institute of Health at Charité – Universitätsmedizin Berlin
Anna-Louisa-Karsch-Str 2
Berlin, 10178
Germany
Phone: 49 30450543049
Email: carina-nina.vorisek@charite.de

Abstract

Background: Resources are increasingly spent on artificial intelligence (AI) solutions for medical applications aiming to improve diagnosis, treatment, and prevention of diseases. While the need for transparency and reduction of bias in data and algorithm development has been addressed in past studies, little is known about the knowledge and perception of bias among AI developers.

Objective: This study's objective was to survey AI specialists in health care to investigate developers' perceptions of bias in AI algorithms for health care applications and their awareness and use of preventative measures.

Methods: A web-based survey was provided in both German and English language, comprising a maximum of 41 questions using branching logic within the REDCap web application. Only the results of participants with experience in the field of medical AI applications and complete questionnaires were included for analysis. Demographic data, technical expertise, and perceptions of fairness, as well as knowledge of biases in AI, were analyzed, and variations among gender, age, and work environment were assessed.

Results: A total of 151 AI specialists completed the web-based survey. The median age was 30 (IQR 26-39) years, and 67% (101/151) of respondents were male. One-third rated their AI development projects as fair (47/151, 31%) or moderately fair (51/151, 34%), 12% (18/151) reported their AI to be barely fair, and 1% (2/151) not fair at all. One participant identifying as diverse rated AI developments as barely fair, and among the 2 undefined gender participants, AI developments were rated as barely fair or moderately fair, respectively. Reasons for biases selected by respondents were lack of fair data (90/132, 68%), guidelines or recommendations (65/132, 49%), or knowledge (60/132, 45%). Half of the respondents worked with image data (83/151, 55%) from 1 center only (76/151, 50%), and 35% (53/151) worked with national data exclusively.

Conclusions: This study shows that the perception of biases in AI overall is moderately fair. Gender minorities did not once rate their AI development as fair or very fair. Therefore, further studies need to focus on minorities and women and their perceptions of AI. The results highlight the need to strengthen knowledge about bias in AI and provide guidelines on preventing biases in AI health care applications.

(J Med Internet Res 2023;25:e41089) doi: 10.2196/41089

KEYWORDS

bias; artificial intelligence; machine learning; deep learning; FAIR data; digital health; health care; online; survey; AI; application; diagnosis; treatment; prevention; disease; age; gender; development; clinical

Introduction

Due to the growing amount of health data combined with the desire to gain ground in precision medicine, artificial intelligence (AI) is advancing at a rapid pace across the health care sector [1]. AI applications in medicine enable the analysis of a wide variety of health data types, ranging from web-based applications using machine learning (ML) to physical applications such as intelligent prostheses [2] and sophisticated robots [3]. ML is a subset of AI using large data inputs and outputs with the goal of recognizing patterns leading to autonomous recommendations or decisions. It can be categorized as follows: unsupervised (ability to find patterns), supervised (based on previously provided labels), and reinforcement learning (sequences of rewards and punishments) [4]. Deep learning (DL), another subset of AI, is a class of ML models using artificial neural networks to learn complex relationships between features and labels operating directly on raw or minimally processed data [5]. In a health care setting, AI models require substantial data with highly comprehensive and, in some cases, longitudinal patient information. However, health data sets generally lack a common structure, format, and standardization [6]. In addition, data are not generally integrated across all health care providers but exist relatively isolated in electronic health care records and are therefore prone to being biased [7].

To derive the greatest use of medical AI applications, algorithms must be fair, meaning key population characteristics that impact algorithm outcomes and target variables are considered in the algorithm. Bias in AI, also called algorithmic bias, can be described as an ML model yielding a systematically wrong outcome [8] because of the differential consideration of certain informational aspects, such as gender, age, or ethnic group, contained in a data set [9]. There are already documented examples of biases in AI applications such as facial recognition, in which algorithms perform poorly with faces of females or Black individuals [10], or natural language processing, in which human-like gender biases occur [8,9,11]. A major problem for algorithms used on health data is the distributional shift, meaning a mismatch between training and test data leading to erroneous predictions. When there is bias in the test data and therefore not correct representations of the general population, known inherent variations in the population are not correctly reflected in the resulting output [12].

In particular, the influence of sociocultural gender and biological sex on health conditions is often ignored in algorithm design and study data. One example is the algorithm predicting acute kidney injury, which was trained on a data set containing only 6% females and hence had lower performance among that demographic subgroup [13]. However, sex- and gender-specific differences have been described in several pathological and physiological processes [14-18]. Even recently, it was shown that only 18% of clinical trials on COVID-19 registered on ClinicalTrials.gov and published in scientific journals reported sex-disaggregated results or subgroup analysis [19]. Despite potential biases, defined methods for bias detection and prevention are not officially mandated within the development of AI applications, despite their existence. Explainable artificial intelligence (XAI) refers to algorithms that meet interpretability and completeness requirements [20]. XAI aims to establish transparency by explaining what decisions led to the creation of the algorithm in addition to its inputs and outcomes, which provide the basis for trusting the algorithm [21]. Methods through which XAI can be established include layer-wise relevance propagation [22] and rationalization [23]. Furthermore, Friedrich et al [24] discussed the role and benefits of statistics, which they see as a natural partner in AI developments, for example, in calculating the sample size but also for bias control. As the existence of biases in AI is a known fact, little is known about the perception and knowledge of AI developers on biases and their prevention. This study set out to answer the following research questions from the AI developers' perspective: (1) How fair are current AI developments? (2) What is preventing fair AI algorithms? and (3) What data are used to train AI algorithms?

Methods

Survey Design

Survey development was initiated after a literature review to establish an understanding of the current state of knowledge on bias in AI algorithms and after consultation with several experts in the fields of bias and AI development who signed on as coauthors. The team of the Berlin Institute of Health (BIH), comprised of medical doctors with expertise in interoperability and digital medicine, drafted all survey questions in collaboration with the partners.

The final survey (version 5) was the result of an iterative process of developing questions that comprised the first 4 survey versions, consulting experts for review and validation in each instance, and making adjustments in the wording and the question design. The first questionnaire consisted of 29 questions without adaptive questioning. The final web-based questionnaire contained 4 routing and completeness variables, 22 base questions, and 15 questions that would appear based on specific answer choices. The final set of questions included 5 new ones compared to the first version, added to gather demographic details (age and workplace) and qualitative insights into the familiarity of respondents with bias prevention measures. All survey questions except for 3 (1 dichotomous question and 2 Likert-scale questions) were designed to collect qualitative data. Overall, the survey that went live had a maximum of 41 questions, using adaptive questioning to reduce the complexity and volume of questions. Single- and multiple-choice questions were included, as were free-text fields for further explanatory comments. Categorical questions included a "not specified" option to select for nonresponse. The questionnaire was developed both in German and English and can be found in Multimedia Appendix 1.

Study data were collected and managed using REDCap (Vanderbilt University) tools hosted at Charité - Universitätsmedizin Berlin [25,26]. On the survey landing page, participants were informed about the survey goals, its length, the target group, and the investigator organizations in both English and German. After choosing the survey language, participants were redirected to the first survey section.

Participation in this anonymous study was voluntary, and no incentives were provided. The first survey page included demographical questions regarding country, institution, gender, age, and work environment. The participant could only proceed to and answer the next section of the questionnaire if they answered that they had experience in AI development; otherwise, the survey ended at this point. The second survey page focused on the type of AI development and medical specialty the participants were involved in. Respondents were able to review and change answers by clicking on the "Previous Page" button shown on the survey screen. The comprehensibility, usability, and technical functionality of the survey were tested by several employees of the inquiring institutions prior to its launch. We used REDCap's internal functionality to check the completeness of each questionnaire page after a participant had submitted it.

Code Availability

All R scripts that were written and used for analysis in this study can be accessed via a dedicated GitHub repository, which is referenced in the data availability statement.

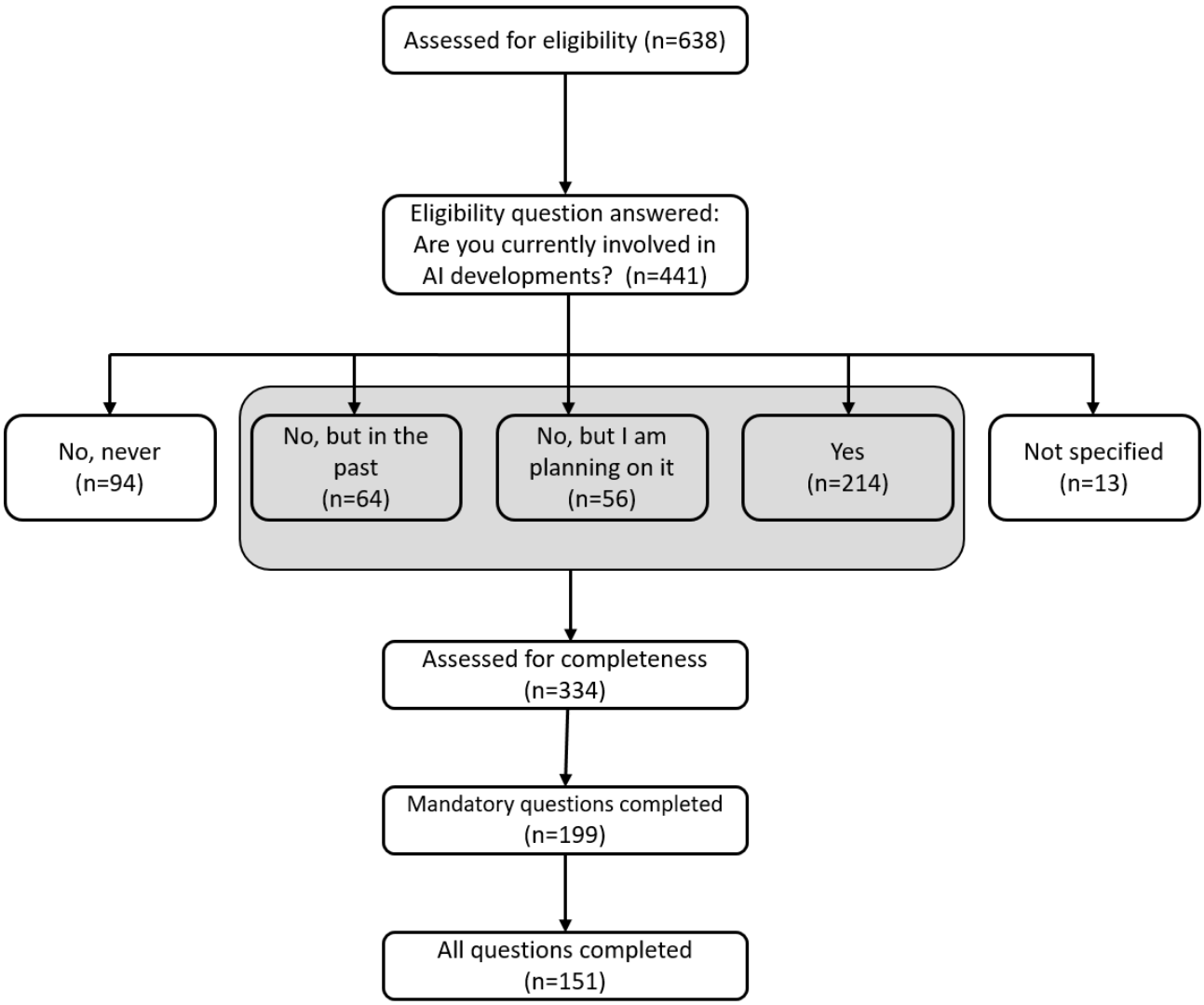
Recruitment

The target group of the survey was participants with experience in developing AI applications in health care. The survey was distributed on the web as an open survey inviting voluntary participation. Announcements were made through the newsletters of universities, other institutions, associations, and organizations in digital health, through mailing lists, as well as by contacting professionals directly via email. In addition, social media accounts (LinkedIn and Twitter) and web magazines were used to recruit participants. An overview of the wording used to announce the survey and invite participation is shown in Multimedia Appendix 2. We provided a dedicated URL leading to the survey via a link. The survey was kept open from August 20, 2021, to November 20, 2021.

Data Exclusion

The primary criteria for inclusion in this study were the responses to the question, "Are you currently involved in AI developments?" Only entirely completed questionnaires from participants with experience in AI development were included. The process of data exclusion is demonstrated in Figure 1.

Figure 1. Flow diagram outlining the number of eligible and responding participants to the survey, as well as the number of participants included in the analysis. AI: artificial intelligence.



Data Analysis

Tables and Likert graphs were used to present the data synthesis results. For categorical demographic variables, simple and relative frequency and proportions were used. Continuous variables are presented in median and IQR unless otherwise stated. Categories for age groups were created based on the distribution of data. For reproducibility purposes, the survey data were downloaded from REDCap directly into an R project, and analysis of the survey data was performed using RStudio (version 1.4.1106; R Studio, PBC).

Ethical Considerations

All participants provided consent and were provided with the survey duration and purpose of the study prior to the survey. No personal information was collected or stored. There were no incentives offered to complete the survey. According to the "Ärztliche Berufsordnung, page 8, §15 Forschung (1)," ethical approval for this study was not requested.

Results

Demographics

A total of 638 participants answered the survey and were assessed for eligibility. Of the 441 participants who answered the eligibility question, 107 were excluded due to a lack of experience in AI development. After further exclusion of 183 incomplete questionnaires, a total of 151 participants involved in AI development completed the survey with a median age of 30 (IQR 26-39) years. The majority of respondents worked in Germany (139/151, 92%), while 2% (3/151) worked in the United States and less than 1% (1/151) in Austria, the Czech Republic, Finland, France, Hungary, the Netherlands, Spain, Switzerland, and Scotland, respectively. Participants received the survey via email distribution (72/151, 48%), personal contact (25/151, 17%), LinkedIn (25/151, 17%), Twitter (11/151, 7%), other options (16/151, 11%), or did not specify the channel of survey reception (2/151, 1%). Details on demographics are found in Table 1 and Table S2 in Multimedia Appendix 2.

Table 1. Demographics of survey participants (N=151).

Characteristics	Participants, n (%)
Gender	
Male	101 (67)
Female	45 (30)
Undefined	2 (1)
Not specified	2 (1)
Diverse	1 (1)
Age (years)	
≤30	69 (46)
30-40	49 (33)
41-50	24 (16)
≥50	9 (6)
Work environment	
Science	104 (69)
Industry	22 (15)
Clinical work	12 (8)
Other	11 (7)
Not specified	2 (1)
Most common medical specialties	
Digital medicine	38 (25)
Radiology	32 (21)
Not specified	29 (19)
Other	22 (15)
Internal medicine	21 (14)
Surgery	14 (9)
Family medicine	14 (9)
Public health	11 (7)
Neurology	10 (7)
Stage of AI[a] project	
Training and optimization	105 (70)
Data acquisition or preprocessing	84 (56)
Identification of AI algorithms	83 (55)
Project planning	82 (54)
Practical testing of AI algorithms	79 (52)
Data annotation	61 (40)
Not specified	7 (5)
Other	6 (4)

[a]AI: artificial intelligence.

AI Experience of Participants

The majority (105/151, 70%) of participants used ML within their AI project, followed by DL (86/151, 57%) and other types of AI (41/151, 27%). Among participants working with ML, 53% (80/151) used supervised ML, 28% (42/151) semisupervised ML, 27% (41/151) unsupervised ML, and 14% (21/151) reinforcement learning. Five percent (7/151) of respondents used other ML techniques and 4% (6/151) did not specify their answer. Participants working with DL used convolutional networks (71/151, 47%), recurrent neural networks (40/151, 26%), autoencoders (30/151, 20%), and other

types of DL (26/151, 17%). Participants used natural language processing (39/151, 26%), clinical decision support (53/151, 35%), image processing (64/151, 42%), computer vision (50/151, 33%), and robotics (16/151, 11%) within their AI developments.

Current Knowledge of Biases in AI

Most respondents (113/151, 75%) had heard of biases before and knew specific use cases, while 20% (30/151) of respondents could not think of concrete examples. Five percent (8/151) had never heard of biases in AI before. When asking respondents where they think biases in AI could possibly occur, the majority voted for societal factors (126/151, 83%), followed by the methodology of algorithms (99/151, 66%), data validation, or data security (119/151, 79%). No respondent answered that none of these options could provide biases, and 12% (18/151) felt that there were other parameters that could lead to biases in AI. Table 2 presents the knowledge distribution in terms of preventive measures to avoid biases in AI.

Table 2. Do you know any of the following preventive measures to avoid bias in AI applications?

Knowledge of preventive measures	Participants, n (%)
XAI[a]	85 (56)
Collecting sociodemographic data	53 (35)
Statistical analysis	95 (63)
Software evaluating fairness in AI[b]	44 (29)
I do not know any of them	25 (17)
Other	3 (2)

[a]XAI: explainable artificial intelligence.

[b]AI: artificial intelligence.

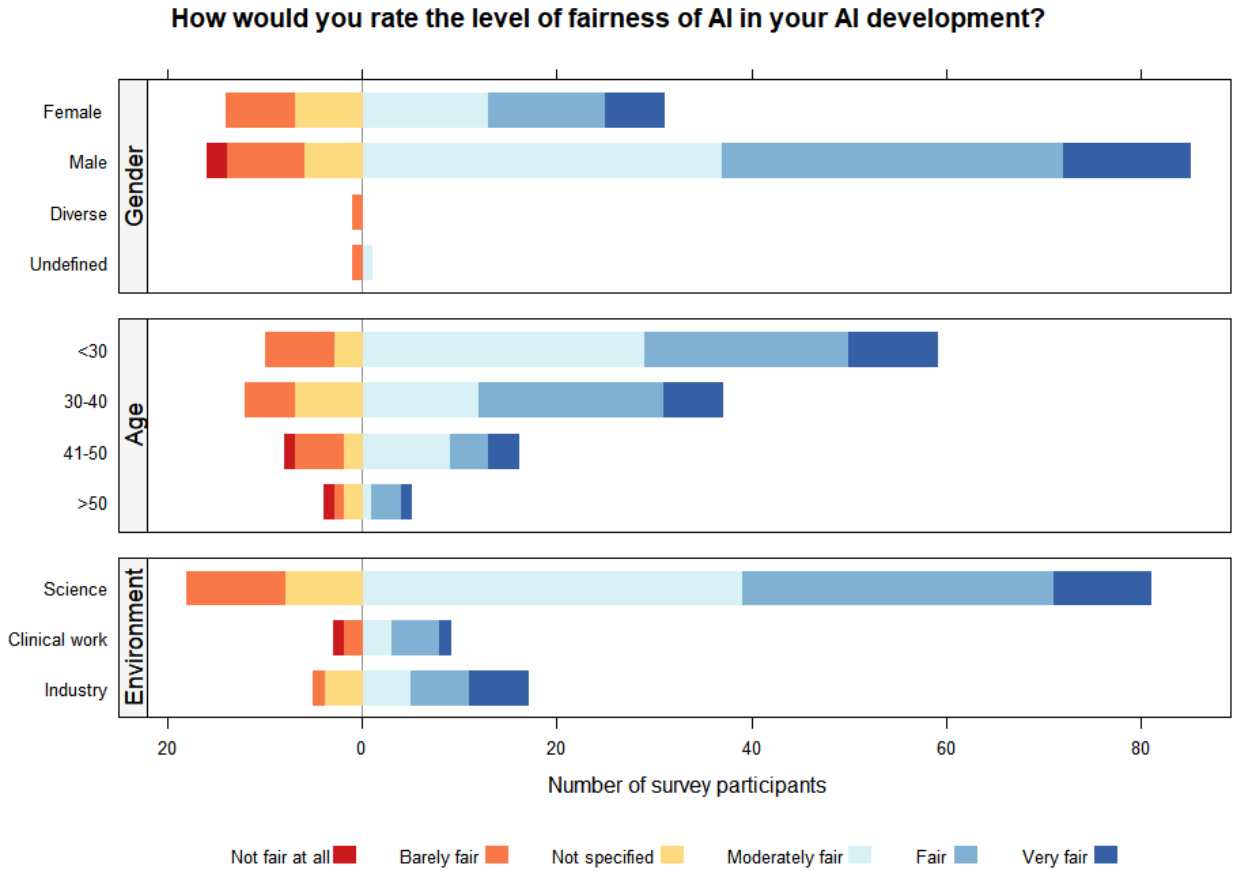
Data Used in AI Projects

Half of the respondents worked with image (83/151, 55%) and text data (77/151, 51%). Audio data were used by 12% (18/151) of respondents, and 16% (24/151) of respondents did not specify the type of data used in their AI projects. Regarding the origin of the data, half of the respondents used data from 1 center (76/151, 50%). Multicenter databases (50/151, 33%), registries (25/151, 17%), and wearables (18/151, 12%) were also used for AI algorithms. Other or unspecified data sources were used by 21% (32/151) and 12% (18/151) of respondents, respectively. Thirty-five percent of respondents (53/151) used national data only, 33% (50/151) used national and international data, and 13% (19/151) worked with international data only.

Prevention of Biases in AI

When asked whether standardized data by using international semantic and syntactic standards such as HL7 FHIR or SNOMED CT could reduce bias in AI, only 25% (37/151) answered “yes,” 44% (66/151) answered “no,” and 32% (48/151) did not specify their answer (Figure 2). Regarding including sociodemographic information in training data for AI algorithms, most of the respondents (100/151, 66%) would collect data on age to prevent biases, followed by biological gender (95/151, 63%) and origin (95/151, 63%). Social gender was chosen by 54% (81/151) of participants, and 6% (9/151) would collect none of the suggested data points. When asked what the participants would use the sociodemographic data for, 62% (93/151) said they would use it for analysis of data, while 38% (57/151) would use the data for AI modeling and 40% (61/151) for data acquisition.

Figure 2. Likert graphs displaying the perception of fair AI among AI developers. The horizontal axis corresponds to the absolute number of survey participants, while the vertical axis details the absolute response counts. The level of fairness was shaded in colors, as detailed in the figure legend. AI: artificial intelligence.



Current Perception of Biases in AI

When the entire cohort was asked how they would rate the level of fairness of AI in their own AI development, one-third rated their development as fair (47/151, 31%) or moderately fair (51/151, 34%). AI developments were rated as very fair by 13% (19/151) of respondents. Twelve percent (18/151) gave a rating for barely fair and 1% (2/151) for not fair at all. Among the 132 (87%) respondents who did not rate their project as very fair, possible reasons preventing fair AI were explored: the majority of respondents selected lack of fair data (90/132, 68%), followed by lack of guidelines or recommendations (65/132, 49%), as well as lack of knowledge (60/132, 45%). Lack of support from superiors, institutions, or other options was answered by 9% (12/132) and 5% (7/132), respectively.

Level of Fairness Perception by Gender, Age Group, and Work Environment

Figure 2 shows the perception of fairness by gender, age group, and work environment. We defined fairness as whether the decision-making process of the algorithm takes sensitive factors like gender, race, and age into account, meaning whether specific groups of people are treated or considered differently from others [27]. There was a significant difference among participants with regard to the work environment (P=.02): 5% (1/22) of respondents working in the industry compared to 10% (10/104) working in science, and 25% (3/12) of respondents working in a clinical setting rated their AI developments as not fair at all or barely fair. There was no significant difference in the level of fairness expressed by respondents in terms of their gender (P=.17) or age group (P=.09). The 3 participants identifying as diverse or undefined rated their AI development as barely fair (1/1) or barely fair (1/2) and moderately fair (1/2), respectively, while male participants had the highest percentage of a very fair perception (13/101, 13%).

Discussion

Principal Findings

Biased algorithms in health care are not news; however, established guidelines, laws, or literature on the practical use of fair AI algorithms in health care are scarce [28]. Therefore, we investigated whether AI developers perceived AI algorithms as fair, as well as their knowledge of preventive measures. We found that one-third of participants rated their AI project as either fair or moderately fair, while 13% (20/151) rated their AI project as barely fair or not fair at all. The main reasons for biases were a lack of fair data, guidelines, and recommendations, as well as knowledge of biases in AI. There was no difference in bias perception among different genders; however, participants did not represent a diverse overall population. The majority of respondents were male, younger than 35 years, and employed in a scientific work environment. Half of the respondents worked with image data from 1 center only, and one-third worked with national data only.

Most AI projects were in the current stage of training and optimization, while only 52% (79/151) performed practical

testing of AI algorithms. This could also introduce bias into our results, as possible AI biases might occur during testing when models perform poorly in data sets that differ from the training data [13,29]. When asked about preventive measures regarding biases in AI, more than half of the participants were aware of methods such as XAI and statistical analysis. Less than half of the participants were aware of collecting sociodemographic data or software evaluating fairness in AI, highlighting the need for education on these methods. Interestingly, 17% (25/151) of participants did not know any of these preventive measures, and 5% of AI developers had never heard of biases in AI before calling for general training to prevent biases in AI.

The data types most commonly used for AI projects were image data, followed by text data. This coincides with the fact that the second most common medical specialty after digital medicine was the field of radiology, in which survey participants worked. In terms of data availability, most data originated from 1 center only in a national setting. One-third used international data in addition to national data, and a small proportion worked with international data only. This might be concerning as the increase in training set size can reduce discrimination within AI [30].

Only one-quarter of participants felt that using international standards could reduce biases, despite interoperability being one of the main parts of the FAIR data principles, meaning findable, accessible, interoperable, and reusable [28,31]. Interoperability describes the ability of systems to exchange data and use the data after these have been received [32]. While there are several levels of data interoperability (semantic, syntactic, organizational, etc) [33], the prevention of bias in AI can be assisted through the use of standard terminologies and ontologies. International standard codes provide an unambiguous meaning to health care concepts such as diagnosis, drug prescriptions, and demographic parameters present in data and can facilitate semantic interoperability [34,35].

Overall, developers participating in this study supported the inclusion of sociodemographic parameters. The 3 characteristics of age, biological sex, and ethnicity were most supported by respondents. The significance of including such data for the purpose of data acquisition and analysis as well as AI modeling should not be underestimated. All categories influence many physiological and biochemical processes and can drive pathogenic and homeostatic phenotypes [36], important variables that AI models aim to predict or model. Social gender would have been collected by more than half of the respondents. However, most respondents would use this sociodemographic data more for analysis than AI modeling.

The fact that lack of fair data was named as the main reason for reported shortcomings in terms of fairness highlights the potential and importance of establishing fair and accessible training data for the development of algorithms. In addition, the lack of usable guidelines or recommendations for establishing fairness in AI was mentioned as the second obstacle to fair AI. This highlights a clear need for guidance on how fair AI can technically be implemented and achieved in AI health care applications. Arguably, such guidance could best be provided by accredited international institutions with strong competencies in the fields of AI and health care, as well as an excellent understanding of the FAIR data principles [31]. Recent publications have presented approaches defining ethical and regulatory boundaries for AI applications in health care, including a governance model [37], a bias evaluation checklist for predictive models used in hospital settings [38], and an overview of ethical principles to be considered in the regulation of bias in health care ML [39]. Furthermore, in July 2018, the World Health Organization established the Focus Group on “Artificial Intelligence for Health” together with the International Telecommunication Union. The Focus Group on “Artificial Intelligence for Health” aims to identify concerns in health care AI regarding data, processes, and algorithms and to develop guidelines while creating a web-based platform and benchmarking tools for health AI [40].

When asking participants whether they felt their AI development was perceived as fair, 13% (19/151) of respondents rated their work as very fair, while approximately 65% (98/151) of participants considered their projects fair or moderately fair. However, the majority of respondents in this survey cohort were male. As the participants were not as diverse when mostly represented by male AI developers, these results could be skewed, and a future aim would be to specifically target opinions by social minorities regarding health data. This is supported by the fact that the 3 participants in our survey who identified as diverse or undefined rated their AI development as barely fair or moderately fair only. Among AI specialists in health care, especially in Germany, one of the leading AI countries, there is an underrepresentation of women, as only 24% of AI specialists are female [41]. The fact that AI specialists are not prone to diversity does not only relate to gender but also race and other underrepresented groups and should be targeted, as including diverse stakeholders was previously reported to be a recommendation for incorporating fairness into AI [29]. A future aim would be to specifically seek out the opinions of social minorities and women regarding fairness in health data AI through a targeted questionnaire that could be distributed using social media and professional networks.

When investigating the perception of fairness among work environments, we found that the majority of survey participants in a scientific work environment rated their AI development only as moderately fair, while the majority from the industry evaluated their application as fair or even very fair. Developers in clinical work were the only ones that rated their algorithms as not fair at all, and they also constituted the highest subgroup that rated algorithms as barely fair. The survey data do not point to a clear explanation for this observation. When it comes to data in health care, the fact that especially hospital information systems were not created to allow for modern analytics and the lack of full integration with other relevant external and internal data systems present a significant challenge [8]. In addition, data generated in experimental, preclinical, or clinical research settings that may serve as the basis for developing algorithms often contain inherent sex biases due to the overrepresentation of male study subjects over females. When working with a limited budget and resources such as time and workforce, ensuring fairness in AI might be neglected in favor of using resources for other purposes due to the lack of importance

attributed to it. A more in-depth follow-up study would aid in investigating this further.

Limitations and Outlook

While the power of this study to draw conclusions might be limited due to the relatively small sample size [42], the survey results provide a first set of insights into the use of fair data in AI development for health care use cases. Future studies will have to focus on addressing a larger AI developer audience in order to reduce response bias [43]. As the majority of participants worked in Germany and were mostly male, the survey may not be generalizable. Furthermore, another limitation is the web-based nature of this survey, which lacks a formal sampling frame. Since the majority of survey questions were designed to collect qualitative data, the determination of reliability and validity of questions using measures such as Cronbach α [44], which require the presence of at least 3 quantitative questions, cannot be performed. Issuing a second, more targeted, and structured survey in a year's time would allow for a follow-up probe into how the field is evolving in industry versus a scientific and clinical setting. The responses from this study have shed new light on developers' awareness of bias in AI and showed that there is a need for education on preventive measures, especially with regard to fair data and the FAIR principles, as well as including sociodemographic factors for training AI algorithms. Guidelines and recommendations are warranted to guarantee fair algorithms that are generalizable to the target population.

Acknowledgments

We would like to thank all survey participants for taking the time to answer the survey and for their valuable contributions to gaining new insights into the fairness of AI (artificial intelligence) development in health care. We also appreciate Alexander Bartschke from the Berlin Institute of Health (BIH) at Charité-Universitätsmedizin Berlin and Andreas Hetey from Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Clinical Trial Office, for their help with the survey tools in REDCap. This work was supported by the BIH Gender Equality Fund.

Data Availability

The survey protocol, data, and analysis scripts can be accessed via a dedicated GitHub repository [45].

Conflicts of Interest

ST is vice chair of HL7 Deutschland e.V. AD is the chairwoman of the Supervisory Board of Dr. Jäschke AG, a member of the Interop Council, and has received lecture/consulting fees from ValueBasedManagedCare GmbH, Janssen-Cilag GmbH, Getinge Deutschland GmbH, EIT Health Germany GmbH, Face-to-face GmbH, Siemens Aktiengesellschaft, gematik GmbH, Medical Congresses & Events e.K., and Pfizer GmbH. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1

Questionnaire.
[PDF File (Adobe PDF File), 685 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Supplementary tables.
[DOCX File , 42 KB-Multimedia Appendix 2]

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
2. Fajardo J, Maldonado G, Cardona D, Ferman V, Rohmer E. Evaluation of user-prosthesis-interfaces for sEMG-based multifunctional prosthetic hands. Sensors (Basel) 2021;21(21):7088 [FREE Full text] [doi: 10.3390/s21217088] [Medline: 34770393]
3. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69S:S36-S40 [doi: 10.1016/j.metabol.2017.01.011] [Medline: 28126242]
4. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020;9(2):14 [FREE Full text] [doi: 10.1167/tvst.9.2.14] [Medline: 32704420]
5. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436-444 [doi: 10.1038/nature14539] [Medline: 26017442]
6. Noorbakhsh-Sabet N, Zand R, Zhang Y, Abedi V. Artificial intelligence transforms the future of health care. Am J Med 2019;132(7):795-801 [FREE Full text] [doi: 10.1016/j.amjmed.2019.01.017] [Medline: 30710543]
7. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform 2015;53:162-173 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.006] [Medline: 25463966]

8. Nelson GS. Bias in artificial intelligence. N C Med J 2019;80(4):220-222 [FREE Full text] [doi: 10.18043/ncm.80.4.220] [Medline: 31278182]
9. Yoon DY, Mansukhani NA, Stubbs VC, Helenowski IB, Woodruff TK, Kibbe MR. Sex bias exists in basic science and translational surgical research. Surgery 2014;156(3):508-516 [doi: 10.1016/j.surg.2014.07.001] [Medline: 25175501]
10. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK. Face recognition performance: role of demographic information. IEEE Trans Inform Forensic Secur 2012;7(6):1789-1801 [doi: 10.1109/tifs.2012.2214212]
11. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science 2017;356(6334):183-186 [FREE Full text] [doi: 10.1126/science.aal4230] [Medline: 28408601]
12. Towards trustable machine learning. Nat Biomed Eng 2018 Oct;2(10):709-710 [doi: 10.1038/s41551-018-0315-x] [Medline: 31015650]
13. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 2019;572(7767):116-119 [FREE Full text] [doi: 10.1038/s41586-019-1390-1] [Medline: 31367026]
14. Zhu C, Boutros PC. Sex differences in cancer genomes: much learned, more unknown. Endocrinology 2021;162(11):bqab170 [FREE Full text] [doi: 10.1210/endocr/bqab170] [Medline: 34402895]
15. The Lancet Neurology. A spotlight on sex differences in neurological disorders. Lancet Neurol 2019;18(4):319 [FREE Full text] [doi: 10.1016/S1474-4422(19)30001-8] [Medline: 30878094]
16. Takahashi T, Ellingson MK, Wong P, Israelow B, Lucas C, Klein J, Yale IMPACT Research Team, et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. Nature 2020;588(7837):315-320 [FREE Full text] [doi: 10.1038/s41586-020-2700-3] [Medline: 32846427]
17. Mauvais-Jarvis F. Gender differences in glucose homeostasis and diabetes. Physiol Behav 2018;187:20-23 [FREE Full text] [doi: 10.1016/j.physbeh.2017.08.016] [Medline: 28843891]
18. Shang D, Wang L, Klionsky DJ, Cheng H, Zhou R. Sex differences in autophagy-mediated diseases: toward precision medicine. Autophagy 2021;17(5):1065-1076 [FREE Full text] [doi: 10.1080/15548627.2020.1752511] [Medline: 32264724]
19. Brady E, Nielsen MW, Andersen JP, Oertelt-Prigione S. Lack of consideration of sex and gender in COVID-19 clinical studies. Nat Commun 2021;12(1):4015 [FREE Full text] [doi: 10.1038/s41467-021-24265-8] [Medline: 34230477]
20. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. arXiv Preprint posted online on February 3, 2019. [FREE Full text] [doi: 10.1109/dsaa.2018.00018]
21. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 2022;77:29-52 [FREE Full text] [doi: 10.1016/j.inffus.2021.07.016] [Medline: 34980946]
22. Montavon G, Binder A, Lapuschkin S, Samek W, Müller KR. Layer-wise relevance propagation: an overview. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR, editors. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer International Publishing; 2019:193-209
23. Ehsan U, Harrison L, Chan L, Riedl MO. Rationalization: a neural machine translation approach to generating natural language explanations. arXiv Preprint posted online on December 19, 2017. [FREE Full text] [doi: 10.48550/arXiv.1702.07826]
24. Friedrich S, Antes G, Behr S, Binder H, Brannath W, Dumpert F, et al. Is there a role for statistics in artificial intelligence? Adv Data Anal Classif 2021;16(4):823-846 [FREE Full text] [doi: 10.1007/s11634-021-00455-6]
25. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
27. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82-115 [doi: 10.1016/j.inffus.2019.12.012]
28. Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. Inf Serv Use 2017;37(1):49-56 [FREE Full text] [doi: 10.3233/isu-170824]
29. Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? JAMA 2018;319(19):1981-1982 [FREE Full text] [doi: 10.1001/jama.2018.2009] [Medline: 29710130]
30. Chen I, Johansson FD, Sontag D. Why is my classifier discriminatory? arXiv Preprint posted online on December 10, 2018. [FREE Full text] [doi: 10.5260/chara.21.2.8]
31. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]

32. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. In: IEEE Std 610. New York, NY: Institute of Electrical and Electronics Engineers; 1991:1-217
33. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR, 3rd Edition. London: Springer Verlag; 2016.
34. Arvanitis TN. Semantic interoperability in healthcare. Stud Health Technol Inform 2014;202:5-8 [Medline: 25000001]
35. Gansel X, Mary M, van Belkum A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. Eur J Clin Microbiol Infect Dis 2019;38(6):1023-1034 [doi: 10.1007/s10096-019-03501-6] [Medline: 30771124]
36. Franconi F, Campesi I. Pharmacogenomics, pharmacokinetics and pharmacodynamics: interaction with biological differences between men and women. Br J Pharmacol 2014;171(3):580-594 [FREE Full text] [doi: 10.1111/bph.12362] [Medline: 23981051]
37. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc 2020;27(3):491-497 [FREE Full text] [doi: 10.1093/jamia/ocz192] [Medline: 31682262]
38. Wang HE, Landers M, Adams R, Subbaswamy A, Kharrazi H, Gaskin DJ, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. J Am Med Inform Assoc 2022;29(8):1323-1333 [FREE Full text] [doi: 10.1093/jamia/ocac065] [Medline: 35579328]
39. McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Zlotnik Shaul R. Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. J Am Med Inform Assoc 2020;27(12):2024-2027 [FREE Full text] [doi: 10.1093/jamia/ocaa085] [Medline: 32585698]
40. Wiegand T, Lee N, Pujari S, Singh M, Xu S, Kuglitsch M, et al. Whitepaper for the ITU/WHO focus group on artificial intelligence for health. International Telecommunication Union. 2019. URL: https://www.itu.int/en/ITU-T/focusgroups/ai4h/Documents/FG-AI4H_Whitepaper.pdf [accessed 2023-03-24]
41. Global gender gap report 2020. World Economic Forum. 2019. URL: https://www3.weforum.org/docs/WEF_GGGR_2020.pdf [accessed 2023-03-24]
42. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Med (Zagreb) 2021;31(1):010502 [FREE Full text] [doi: 10.11613/BM.2021.010502] [Medline: 33380887]
43. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf 2019;28(3):231-237 [FREE Full text] [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
44. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011;2:53-55 [FREE Full text] [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
45. Analysis on survey bias in AI. GitHub. 2022. URL: https://github.com/BIH-CEI/BiasAI [accessed 2023-03-24]

Abbreviations

AI: artificial intelligence
BIH: Berlin Institute of Health
DL: deep learning
ML: machine learning
XAI: explainable artificial intelligence

Edited by A Mavragani; submitted 22.07.22; peer-reviewed by A Burmann, O Sverdlov; comments to author 11.10.22; revised version received 11.11.22; accepted 20.04.23; published 22.06.23

Please cite as:
Vorisek CN, Stellmach C, Mayer PJ, Klopfenstein SAI, Bures DM, Diehl A, Henningsen M, Ritter K, Thun S
Artificial Intelligence Bias in Health Care: Web-Based Survey
J Med Internet Res 2023;25:e41089
URL: https://www.jmir.org/2023/1/e41089
doi: 10.2196/41089
PMID: