Journal of Biomed Research

Kleinberg et al, J Biomed Res. 2022;3(1):42-47.

Reseach Article

Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare

Giona Kleinberg[1], Michael J Diaz[2], Sai Batchu[3], Brandon Lucke-Wold, MD, PhD[4*]

[1]Northeastern University, Boston, MA, United States

[2]University of Florida, College of Medicine, Gainesville, FL, United States

[3]Montville, NJ, United States

[4]Department of Neurosurgery, University of Florida, Gainesville, FL, United States

*Author for correspondence:

Email: Brandon.Lucke-Wold@neurosurgery.ufl.edu

Received date: September 27, 2022
Accepted date: October 11, 2022

Citation: Kleinberg G, Diaz MJ, Batchu S, Lucke-Wold B. Racial underrepresentation in dermatological datasets leads to biased machine learning models and inequitable healthcare. J Biomed Res. 2022;3(1):42-47.

Abstract

Objective: Clinical applications of machine learning are promising as a tool to improve patient outcomes through assisting diagnoses, treatment, and analyzing risk factors for screening. Possible clinical applications are especially prominent in dermatology as many diseases and conditions present visually. This allows a machine learning model to analyze and diagnose conditions using patient images and data from electronic health records (EHRs) after training on clinical datasets but could also introduce bias. Despite promising applications, artificial intelligence has the capacity to exacerbate existing demographic disparities in healthcare if models are trained on biased datasets.

Methods: Through systematic literature review of available literature, we highlight the extent of bias present in clinical datasets as well as the implications it could have on healthcare if not addressed.

Results: We find the implications are worsened in dermatological models. Despite the severity and complexity of melanoma and other dermatological diseases as well as differing disease presentations based on skin-color, many imaging datasets underrepresent certain demographic groups causing machine learning models to train on images of primarily fair-skinned individuals leaving minorities behind.

Conclusion: In order to address this disparity, research first needs to be done investigating the extent of the bias present and the implications it may have on equitable healthcare.

Keywords: Deep learning, Artificial intelligence, Clinical technology, Health disparity

Introduction

Modern advances in artificial intelligence are phenomenal as trained machine learning models are learning to solve humanities greatest problems. In addition to the many technical problems machine learning models are attempting to solve, there is a plethora of research on the possible clinical applications of machine learning [1-5]. Prediction models can be applied to many clinical problems and can function as promising tools for diagnosis, risk analysis, and patient treatment [6-10]. Clinical applications of machine learning are no longer a goal for the future. Recently, artificial intelligence has been successfully implemented in many studies to assist clinical outcomes in the COVID-19 pandemic and work on other pressing medical issues [11-14]. The advantages supervised machine learning can supply are tremendous; improved patient outcomes due to automated screening, assistance for medical providers, and greater efficiency are just a few of the possible advantages using such technology can allow [15,16]. Through training a model on clinically relevant datasets and electronic health records some supervised machine learning models have even performed better than licensed physicians [8,17].

Along with the prominent advantages of machine learning, there are many associated risks with using artificial intelligence for clinical applications [18]. One of the most prominent concerns for the use of machine learning is a biased impact on healthcare. Universally, healthcare systems struggle in order to minimize racial, ethnic, and other demographically based disparities [19,20]. The field of dermatology is no exception to exhibiting such disparities [21]. Due to prominent differences in skin color among patients, dermatological machine learning models may present a high risk of worsening

healthcare disparities due to underrepresentation in training datasets. Dermatological models that learn from biased datasets risk biased performance that may leave minorities behind. In order to inform future artificial intelligence implementation by providers and researchers, it is necessary to investigate and address the extent that machine learning may exacerbate these racial and other demographic disparities in dermatology [22]. This review explores the extent, causes, possible solutions, and overall literature coverage on bias in dermatological machine learning models.

Methods

We searched and selected using the PubMed Advanced Search Builder and MeSH keyword queries. In order to investigate the extent and relation between artificial intelligence and biased clinical applications, we used the advanced keyword query "machine learning" AND "bias" AND "healthcare" yielding 112 results of which 26 were selected for further review. In order to refine the search to literature pertaining to bias specifically in dermatological applications of machine learning, we also searched for "dermatology" AND "machine learning" AND "bias" and obtained 11 results of which 6 were selected for review. The advanced keyword query "dermatology" AND "machine learning" AND "diversity" was also used to investigate underrepresentation and overrepresentation of dermatological datasets and models. The search yielded 7 additional results of which 2 were selected for further review. Lastly, in order to gain insight into the effect that the bias such underrepresentation and overrepresentation may introduce into artificial intelligence models during model training specifically, we searched for "machine learning training" AND "diversity" and "bias" which output 9 results of which 4 were selected for further review.

Strict inclusion criteria included clinical trials (including randomized control trials), meta-analyses, review articles, and clinical artificial intelligence papers. Papers found through the systematic literature search were further assessed for relevance by title and abstract. In order to conduct a focused analysis, only papers focusing on clinical applications of machine learning or the demographic disparity inherent in their use were included and texts based solely on machine learning development outside a clinical setting were therefore excluded. All articles selected were written in English in order to mitigate errors due to mistranslations. A high emphasis was placed on selecting articles in reputable journals due to the quality control inherent in the peer-review process. Some reputable and widely used websites were also included due to their relevance however non-peer-reviewed sources without a strong reputation were excluded. Reputability was primarily determined by the presence of a peer-review process. For sources without a peer review process, reputability was determined by high web traffic and citation count shown in other scientific literature. Eligibility and reputability determinations were discussed, and disagreements were resolved openly by all authors. The search was conducted from January 10, 2022, to January 15, 2022.

Results

Bias in clinical datasets

There is a great amount of literature on the bias of healthcare datasets [7,15,23,24]. Recently, a large portion of the literature has focused on biases in gender leading to greater acknowledgement of the disparity present [25]. Despite this progress, bias in clinical datasets extends to many other areas in addition to gender and still affects a sobering majority of machine learning models [6,12,14]. Risk of bias analysis (RoB) is a statistic computed based on the features of a model and the dataset it is trained on in order to quantify the possible bias present [26]. According to a RoB analysis of machine learning prediction models for COVID-19 diagnosis or prognosis by Adamidi et al, 97% (98/101) of the models included were designated as having a high overall risk of bias. All of the remaining three models were designated as having an unclear risk of bias and no models were designated as having a low risk of bias [11]. This alarmingly high percent of models at high risk for bias is very significant and warrants a search for the cause. A further breakdown of this RoB analysis shows that 30% (30/101) of these studies had a high risk of bias and 21% (21/101) had an unclear risk of bias when looking at participants specifically [11]. The high percentage of risk of bias in features related to participants showcases that the overall high risk of bias in these prediction models is possible to be heavily due to the representation of participants in the training datasets. Other reviews of models further corroborate these consequential results. Another RoB analysis for similar prediction models found that 100% (n=11) of included models were designated as having a high overall risk of bias with 18% (2/11) of studies having an unclear risk of bias in regards to participants [13].

Perhaps unsurprisingly, the large prevalence of biased datasets translates to dermatological datasets as well [27]. In a RoB analysis by Dick et al. 97% (128/132) of melanoma prediction models were designated as having a high risk of bias in at least one category with 44% (58/132) having a high risk of bias in at least two categories [28]. Similar to non-dermatological models, the high risk of bias in many of these models is likely due to under and over representation of various groups in the datasets the models are trained on as evidenced by the high risk of bias in features associated with participants. For one example, the LFW dataset which is a dataset used as a top tier benchmark for face recognition was estimated to consist of 77.5% (10258/13233) male faces and 83.5% (11045/13233) White faces [7,29]. Similarly for dermatological training datasets with a majority of the images being of White patients, models trained on the datasets are highly likely to be biased towards successful prediction on White patients [30].

Absence of demographical data

Although many of these models report high overall prediction success, the breakdown of successful predictions by patient race is not shown for many dermatological datasets [7]. This is another pertinent problem as high-performance metrics and successful prediction rates may misrepresent model accuracy on specific patient groups. If models are deemed successful enough for clinical use due to overall performance metrics, these tools may only improve clinical outcomes for certain racial groups and leave minorities behind or even decrease clinical accuracy due to misinforming providers [7,12,13,17].

Furthermore, many databases and prediction model studies do not report demographic information entirely. In the nationalized health databases of countries such as France or Canada patient ethnicity or race is not reported at all [17]. According to Gupta and Kataraya, a large amount of training data is sourced from social media in order to create clinical datasets. Social media contains great amounts of data leading to large training sets however race, ethnicity, age, gender, or other demographic information is not always explicitly supplied or accurate resulting in datasets with no

available information on representation within the dataset [3]. This is a large problem as underrepresentation or overrepresentation of groups within the dataset cannot be easily identified masking the possible risk of bias.

As a result, the complete extent of racial underrepresentation in dermatological datasets is still unknown. This is partially due to an absence of disclosure of the skin types appearing in the source images in many dermatological datasets specifically. In a systematic review by Guo et al, only 8.82% (12/136) of studies disclosed the race or ethnicity of participants in the source images of the dermatological datasets and only 4.41% (6/136) of the studies disclosed information on skin type [31]. It is very pertinent to identify the overall risk of bias in these datasets and models in order to address the disparities present in existing machine learning approaches [32].

The large prevalence of non-communication regarding race and ethnicity in dermatological datasets could possibly be due to the extreme underrepresentation present [27]. Out of the same 136 studies, only 2 studies included Hispanic individuals, only 1 study explicitly included Black patients and only 1 study explicitly included American Indian or Alaska Native patients [31]. A further analysis of these studies showcased that participant images used to train models were collected primarily in the United States and Europe as well as Australia and various East Asian countries with only 1.47% (2/136) of studies including individuals from South America or Africa [31]. Guo et al also draw attention to many specific popular image repositories that showcase the same underrepresentation; one example being the International Skin Imaging Collaboration whose data is collected predominantly from Europe and Australia with populations of primarily light-skinned individuals [31,33].

Black box algorithms

Similar to the unquantifiable risk of bias in clinical datasets due to a lack of demographical data, the risk of bias introduced through training machine learning models is also difficult to discern [16,17]. As modern artificial intelligence research improves, machine learning models are increasing in complexity and making predictions through the use of features and connections in ways unclear to even the developers of such algorithms [34]. It is important to recognize the inherent risk in using these "black box" algorithms that cannot be completely understood. Kelly et al. highlights the importance of transparent and trustworthy model decision-making as clinical settings require explainable and methodological approaches [16]. Unfortunately, there is an inverse correlation between model performance and model transparency as the best models are usually the most complex. Kelly et al. advise that much greater caution or information is needed in order to use these 'black box' algorithms in a clinical setting [16].

Discussion

Data collection

In order to best address the large prevalence of bias in models, it is helpful to understand how such biases arrive in the datasets they are trained on [24]. One of the most prominent sources of bias is the method of collection used to obtain data for a training set. Data collection can take many forms that can all introduce bias in different ways. Randomized control trials (RCTs) attempt to be unbiased but often have inclusion and exclusion criteria that dramatically decrease the representativeness of the data they obtain [32]. In one example RCT to further asthma treatment, 94% of adults with asthma would not meet the inclusion criteria for the study [17]. Datasets using electronic health records such as the MIMIC-III dataset primarily include data from those who visit their respective intensive care units or emergency departments, but this also introduces biases as mostly White people have access to these healthcare resources while Black or Hispanic people are less likely to receive care at these locations [17]. Underrepresentation of many other groups such as undocumented immigrants and low-income nationals is also partially due to their inaccessibility to the sources of data collection. Racial disparities in these groups are then transferred to healthcare datasets as the datasets are created with the same disparate proportions [17]. Another consideration is data variability. Despite dermatological image collection being relatively easy compared to obtaining other medical images, dermatology images are widely varied and the least standardized [35]. When training models on dermatological datasets, functionality needs to be added to address to large image variety especially due to demographic factors such as skin color [36].

Missing data

Bias can be introduced even after data collection however as even parsing through data can introduce bias. In electronic health record datasets, many incomplete EHRs lower the training size of the dataset. To fix this, many studies will filter the dataset for only "complete" EHRs. According to a study by Weber et al. even this simple filtering to manage missing data introduced a bias towards older female patients [37]. If such a simple filter can introduce gender bias, it is pertinent to investigate the many similar filtering and data wrangling techniques that are common in data science to identify and address all the biases they may incur using recent research advances [38].

Duplicate data

Similar to missing data, duplicate data can also introduce bias into a dataset [11,39]. In large EHR datasets that may be compiled from multiple sources or have multiple patient encounters that are not indexed together, duplicates of various patient EHRs can be mistakenly added into the same training set. This can cause greater inconsistencies in representation as duplicate EHRs can cause greater percentages of various groups. There are many emerging solutions being developed to handle duplication bias such as fold-stratified cross validation, but further research still needs to be done in order to ensure these solutions do not introduce a different bias themselves [39].

Synthetic data

In order to avoid the many possible sources of bias when collecting clinical data, research has been done in order to develop methods of generating data. One alternative method to training prediction models on real data is to use synthetic data (i.e., a large created dataset based off of real data in order to increase the amount of training material for a model). Synthetic data can be useful when training models, however, according to Bhanot et al. synthetic datasets are similarly biased to real datasets. Upon analysis, synthetic versions of three popular clinical datasets all showed considerable bias. The MIMIC-III dataset showed overrepresentation of Whites and Asians and underrepresentation of Blacks. The ATUS dataset which tracks average sleep time of Americans was revealed to show that those 75 years or older or male were greatly underrepresented. In

addition, the autism spectrum disorder (ASD) dataset was revealed to greatly overrepresent Whites and underrepresented Asians [12].

Implications of biased datasets

The most impactful consequence of biased datasets is their effect on a model's diagnoses or other outputs [7,12,13,40]. According to Mpanya et al, the role demographics plays in diagnosis should not be neglected. Many conditions have risk factors and presentations dependent on patient demographics. For example, those in high income countries, such as Europe and the United States, primarily suffer heart failure due to ischemic heart disease while in lower income countries such as sub-Saharan Africa, the predominant cause of heart failure is hypertension [4]. Clinical models trained on an under-representative dataset are more likely to have lower success rates on the underrepresented groups as they will not be sufficiently trained on these valuable connections [41]. Consequently, without a way to distinguish between different races or ethnicities, a model will attempt to diagnose utilizing demographic risk factors and other connections of the majority and will therefore not be as affective on underrepresented groups.

This is especially true when considering dermatological datasets [42]. Skin-lesions, rashes and other dermatological conditions present with great differences based on skin-color, or other demographic factors in source images. Unless models are trained on a representative variety of skin-types, ethnicities, and other factors, they will be unable to predict on underrepresented groups accurately which will further healthcare disparities [27]. In one dataset with over 80% of the training images being of light-skinned individuals, prediction models trained on the dataset could not identify those with skin of color [14]. Another occurrence showcased a facial recognition model trained on a biased dataset that incorrectly classified 28 members of the US Congress as criminals as well as incorrectly classified 40% as the congress members as persons of color when only 20% were [14]. When testing a neural network trained on almost 20,000 images of skin lesions on a second dataset that the model did not train on, Han et al. found the model only correctly diagnosed 29% (29/100) of the lesions correctly indicating that dermatological models can be biased heavily for the dataset and types of images they are trained on [43].

Implications on melanoma

The disparity these inaccuracies cause is increased in dermatological cases as patients with skin of color are already more likely to present with more complex dermatological diseases and have worse survival rates than Whites [16,44,45]. Ideally, machine learning can be used to mitigate these worse survival rates through use as a diagnostic aid. Machine learning, deep learning and artificial intelligence have been used as a tool in cases of pigmentary skin lesions and malignancies, psoriasis, acne, allergic contact dermatitis, autoimmune disorders, and ulcer cases [42]. In addition to these, arguably the most pervasive dermatological area to consider the implications of machine learning for is melanoma or skin cancer cases. The great occurrence and severity of melanoma as well as the outcome improvement from early diagnosis increase the need for accurate machine learning models that can detect skin cancer early on [28,31]. In order to achieve models that can be relied upon, biases must be identified and removed or mitigated sufficiently so that all groups have access to possibly improved patient outcomes.

Conclusion

According to Pot et al, it is difficult to create completely unbiased datasets and models, however as researchers and providers, we should strive to mitigate bias as much as possible moving forward [23]. Machine learning models must be constantly re-training in order to reflect changing disease patterns in order to have the greatest effect [8,16,34,46]. In order to do this, the extent of bias present first needs to be identified to prevent further bias. Negligence of bias in training datasets and prediction models should not be acceptable [47]. As a minimum, racial, ethnic, and other demographic information should be disclosed accompanying all clinical machine learning studies [48,49]. Due to the possibly great negative implications of bias, those creating training datasets and prediction models should take responsibility for ensuring that they are not contributing to increasing disparities in healthcare [9,47]. Furthermore, adding more diverse representation to existing datasets and then retraining models may mitigate bias however a great emphasis should be placed on the development of reduced bias datasets and models rather than attempting to improve existing models [50].

Additionally, according to Dick et al., there is a high likelihood that progress in using machine learning for melanoma diagnosis is hindered due to dermatologists feeling threatened by the technology [28]. The embrace of new machine learning technology by the physicians directly involved with the application of such technology facilitates a way to detect disparities directly in their use. The benefits of using machine learning and artificial intelligence to augment physician diagnoses and dermatology can better be achieved if dermatologists and other healthcare providers viewed the new technology as a tool instead of as competition for their jobs [42]. Consequently, physicians should remain in control of the ultimate diagnosis in order to embrace and work with new advances in artificial intelligence as well as keep them in check [15,24,34,49]. Despite the great risk for bias when considering dermatological imaging due to different skin colors, we have identified a significant gap in the literature focusing on addressing such bias in dermatology. More research should be done to identify the extent of bias present in dermatological models and datasets specifically in order to begin addressing the disparity present.

Finally, the value of successful applications of clinical models is difficult to overstate. By acknowledging, identifying, and removing the widespread racial biases in dermatological imaging datasets and corresponding machine learning models, the benefits of artificial intelligence can be realized by those of any demographic background and the large disparities in healthcare can be further addressed.

Acknowledgments

The authors declare no additional acknowledgements.

Declaration of Interest

The authors did not receive support from any organization for the submitted work and have no competing interests to declare that are relevant to the content of this article. This manuscript has no associated data to be made available.

References

1. Chowdhury M, Cervantes EG, Chan WY, Seitz DP. Use of Machine Learning and Artificial Intelligence Methods in Geriatric Mental Health Research Involving Electronic Health Record or

Administrative Claims Data: A Systematic Review. Frontiers in Psychiatry. 2021;12.

2. Farook TH, Jamayet NB, Abdullah JY, Alam MK. Machine learning and intelligent diagnostics in dental and orofacial pain management: A systematic review. Pain Research and Management. 2021 Apr 24;2021.
3. Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: a systematic review. Journal of Biomedical Informatics. 2020 Aug 1;108:103500.
4. Mpanya D, Celik T, Klug E, Ntsinjana H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. IJC Heart & Vasculature. 2021 Jun 1;34:100773.
5. Wu JH, Liu TA, Hsu WT, Ho JH, Lee CC. Performance and limitation of machine learning algorithms for diabetic retinopathy screening: meta-analysis. Journal of Medical Internet Research. 2021 Jul 5;23(7):e23863.
6. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value in Health. 2015 Mar 1;18(2):137-40.
7. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. Frontiers in Digital Health. 2021 Jun 25;3:65.
8. Lee S, Mohr NM, Street WN, Nadkarni P. Machine learning in relation to emergency medicine clinical and operational scenarios: an overview. Western Journal of Emergency Medicine. 2019 Mar;20(2):219.
9. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, and Government-University-Industry Research Roundtable. 2018. Artificial Intelligence and Machine Learning to Accelerate Translational Research: Proceedings of a Workshop—in Brief. The National Academies Collection: Reports Funded by National Institutes of Health. Washington (DC): National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK513721/.
10. Wissel BD, Greiner HM, Glauser TA, Mangano FT, Santel D, Pestian JP, et al. Investigation of bias in an epilepsy machine learning algorithm trained on physician notes. Epilepsia. 2019 Sep;60(9):e93-8.
11. Adamidi ES, Mitsis K, Nikita KS. Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review. Computational and Structural Biotechnology Journal. 2021 Jan 1;19:2833-50.
12. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. Entropy. 2021 Sep 4;23(9):1165.
13. Chee ML, Ong ME, Siddiqui FJ, Zhang Z, Lim SL, Ho AF, et al. Artificial intelligence applications for COVID-19 in intensive care and emergency settings: a systematic review. International Journal of Environmental Research and Public Health. 2021 Jan;18(9):4749.
14. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. British Journal of Anaesthesia. 2020 Dec 1;125(6):843-6.
15. Arora A. Conceptualising artificial intelligence as a digital healthcare innovation: an introductory review. Medical Devices (Auckland, NZ). 2020;13:223.
16. Kelly CJ, Karthikesalingam A. Suleyman Met al. et al. Key challenges for delivering clinical impact with Artificial Intelligence. BMC Med. 2019;17:1-9.
17. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annual Review of Biomedical Data Science. 2021 Jul 20;4:123-44.
18. Crawford K. Artificial intelligence's white guy problem. The New York Times. 2016 Jun 25;25(06).
19. "2021 National Healthcare Quality and Disparities Report." 2022. Accessed January 15. https://www.ahrq.gov/research/findings/nhqrdr/nhqdr21/index.html
20. Betancourt JR, Tan-McGrory A, Flores E, López D. Racial and ethnic disparities in radiology: a call to action. Journal of the American College of Radiology. 2019 Apr 1;16(4):547-53.
21. Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, Correa R, et al. Reading Race: AI Recognises Patient's Racial Identity In Medical Images. arXiv Preprint arXiv:2107.10356. 2021 Jul 21.
22. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019 Mar 12;364.
23. Pot M, Kieusseyan N, Prainsack B. Not all biases are bad: equitable and inequitable biases in machine learning and radiology. Insights Into Imaging. 2021 Dec;12(1):1-0.
24. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care?. AMA Journal of Ethics. 2019 Feb 1;21(2):167-79.
25. Lee MS, Guo LN, Nambudiri VE. Towards gender equity in artificial intelligence and machine learning applications in dermatology. Journal of the American Medical Informatics Association. 2022 Feb;29(2):400-3.
26. Farrah K, Young K, Tunis MC, Zhao L. Risk of bias tools in systematic reviews of health interventions: an analysis of PROSPERO-registered protocols. Systematic Reviews. 2019 Dec;8(1):1-9.
27. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. Journal of Dermatological Treatment. 2020 Jul 3;31(5):496-510.
28. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. JAMA Dermatology. 2019 Nov 1;155(11):1291-9.
29. Han H, Jain AK. Age, gender and race estimation from unconstrained face images. Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5). 2014 Jul;87:27.
30. Ebede T, Papier A. Disparities in dermatology educational resources. Journal of the American Academy of Dermatology. 2006 Oct 1;55(4):687-90.
31. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. J Am Acad Dermatol. 2022 Jul;87(1):157-159.
32. Marshall IJ, Kuiper J, Wallace BC. Automating risk of bias assessment for clinical trials. Inproceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. 2014 Sep 20;88-95.
33. "ISDIS." 2022. Accessed January 14. https://isdis.net/.
34. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial Intelligence, Bias and Clinical Safety. BMJ Qual Saf 28 (3): 231–237.

35. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. Journal of Investigative Dermatology. 2018 Oct 1;138(10):2108-10.

36. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality?. The Journal of Investigative Dermatology. 2018 Oct;138(10):2277.

37. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data". Journal of the American Medical Informatics Association. 2017 Nov 1;24(6):1134-41.

38. Nijman SW, Leeuwenberg AM, Beekers I, Verkouter I, Jacobs JJ, Bots ML, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. Journal of Clinical Epidemiology. 2022 Feb 1;142:218-29.

39. Bey R, Goussault R, Grolleau F, Benchoufi M, Porcher R. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. Journal of the American Medical Informatics Association. 2020 Aug 1;27(8):1244-51.

40. Wiens J, Price WN, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nature Medicine. 2020 Jan;26(1):25-6.

41. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Academic Pathology. 2019 Jan 1;6:2374289519873088.

42. Pai VV, Pai RB. Artificial intelligence in dermatology and healthcare: An overview. Indian Journal of Dermatology, Venereology and Leprology. 2021 Jun 30;87(4):457-67.

43. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. Journal of Investigative Dermatology. 2018 Jul 1;138(7):1529-38.

44. Cormier JN, Xing Y, Ding M, Lee JE, Mansfield PF, Gershenwald JE, et al. Ethnic differences among patients with cutaneous melanoma. Archives of Internal Medicine. 2006 Sep 25;166(17):1907-14.

45. Ward-Peterson M, Acuña JM, Alkhalifah MK, Nasiri AM, Al-Akeel ES, Alkhaldi TM, et al. Association between race/ethnicity and survival of melanoma patients in the United States over 3 decades: a secondary analysis of SEER data. Medicine. 2016 Apr;95(17).

46. Nath V, Yang D, Landman BA, Xu D, Roth HR. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. IEEE Transactions on Medical Imaging. 2020 Dec 29;40(10):2534-47.

47. Shah H. Algorithmic accountability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018 Sep 13;376(2128):20170362.

48. Noseworthy PA, Attia ZI, Brewer LC, Hayes SN, Yao X, Kapa S, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. Circulation: Arrhythmia and Electrophysiology. 2020 Mar;13(3):e007988.

49. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. Journal of Public Health Policy. 2021 Nov 22:1-0.

50. Jiang H, Nachum O. Identifying and correcting label bias in machine learning. InInternational Conference on Artificial Intelligence and Statistics. 2020 Jun 3 (pp. 702-712). PMLR.