Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study

Mike Schaekermann,[a,g,*] Terry Spitz,[a,g] Malcolm Pyles,[b,c,g] Heather Cole-Lewis,[a] Ellery Wulczyn,[a] Stephen R. Pfohl,[a] Donald Martin, Jr.[a] Ronnachai Jaroensri,[a] Geoff Keeling,[a] Yuan Liu,[a] Stephanie Farquhar,[a] Qinghan Xue,[a] Jenna Lester,[b,d] Cían Hughes,[a] Patricia Strachan,[a] Fraser Tan,[a] Peggy Bui,[a] Craig H. Mermel,[a,e] Lily H. Peng,[a,f] Yossi Matias,[a] Greg S. Corrado,[a] Dale R. Webster,[a] Sunny Virmani,[a] Christopher Semturs,[a] Yun Liu,[a,***] Ivor Horn,[a,h,**] and Po-Hsuan Cameron Chen[a,h,****]

[a]Google Health, Mountain View, CA, USA
[b]Advanced Clinical, Deerfield, IL, USA
[c]Department of Dermatology, Cleveland Clinic, Cleveland, OH, USA
[d]Department of Dermatology, University of California, San Francisco, CA, USA

eClinicalMedicine
2024;70: 102479
Published Online 14 March 2024
https://doi.org/10.1016/j.eclinm.2024.102479

Summary

Background Artificial intelligence (AI) has repeatedly been shown to encode historical inequities in healthcare. We aimed to develop a framework to quantitatively assess the performance equity of health AI technologies and to illustrate its utility via a case study.

Methods Here, we propose a methodology to assess whether health AI technologies prioritise performance for patient populations experiencing worse outcomes, that is complementary to existing fairness metrics. We developed the Health Equity Assessment of machine Learning performance (HEAL) framework designed to quantitatively assess the performance equity of health AI technologies via a four-step interdisciplinary process to understand and quantify domain-specific criteria, and the resulting HEAL metric. As an illustrative case study (analysis conducted between October 2022 and January 2023), we applied the HEAL framework to a dermatology AI model. A set of 5420 teledermatology cases (store-and-forward cases from patients of 20 years or older, submitted from primary care providers in the USA and skin cancer clinics in Australia), enriched for diversity in age, sex and race/ethnicity, was used to retrospectively evaluate the AI model's HEAL metric, defined as the likelihood that the AI model performs better for subpopulations with worse average health outcomes as compared to others. The likelihood that AI performance was anticorrelated to pre-existing health outcomes was estimated using bootstrap methods as the probability that the negated Spearman's rank correlation coefficient (i.e., "R") was greater than zero. Positive values of R suggest that subpopulations with poorer health outcomes have better AI model performance. Thus, the HEAL metric, defined as p (R >0), measures how likely the AI technology is to prioritise performance for subpopulations with worse average health outcomes as compared to others (presented as a percentage below). Health outcomes were quantified as disability-adjusted life years (DALYs) when grouping by sex and age, and years of life lost (YLLs) when grouping by race/ethnicity. AI performance was measured as top-3 agreement with the reference diagnosis from a panel of 3 dermatologists per case.

Findings Across all dermatologic conditions, the HEAL metric was 80.5% for prioritizing AI performance of racial/ethnic subpopulations based on YLLs, and 92.1% and 0.0% respectively for prioritizing AI performance of sex and age subpopulations based on DALYs. Certain dermatologic conditions were significantly associated with greater AI model performance compared to a reference category of less common conditions. For skin cancer conditions, the HEAL metric was 73.8% for prioritizing AI performance of age subpopulations based on DALYs.

Interpretation Analysis using the proposed HEAL framework showed that the dermatology AI model prioritised performance for race/ethnicity, sex (all conditions) and age (cancer conditions) subpopulations with respect to pre-existing health disparities. More work is needed to investigate ways of promoting equitable AI performance across

*Corresponding author. Google Health, 1600 Amphitheatre Pkwy, Mountain View, CA, 94043, USA.
**Corresponding author. Google Health, 1600 Amphitheatre Pkwy, Mountain View, CA, 94043, USA.
***Corresponding author. Google Health, 1600 Amphitheatre Pkwy, Mountain View, CA, 94043, USA.
****Corresponding author. Google Health, 1600 Amphitheatre Pkwy, Mountain View, CA, 94043, USA.
E-mail addresses: mikeshake@google.com (M. Schaekermann), ivorh@google.com (I. Horn), liuyun@google.com (Y. Liu), cameron.ph.chen@gmail.com (P.-H. Cameron Chen).
[e]Current affiliation: Precision Neuroscience Corporation, Mountain View, CA, USA.
[f]Current affiliation: Verily Life Sciences, South San Francisco, CA, USA.
[g]Co-first authors.
[h]Co-last authors.

age for non-cancer conditions and to better understand how AI models can contribute towards improving equity in health outcomes.

Funding Google LLC.

Keywords: Artificial intelligence; Machine learning; Health equity; Dermatology

Research in context

Evidence before this study

We searched the PubMed database from Jan 1, 2018 until Aug 31, 2022, for articles using the following search terms: 'health equity' AND ('machine learning' OR 'artificial intelligence'). In addition, we searched relevant computer science journals and conference proceedings using similar search terms. The retrieved articles were analysed with respect to the proposed framework components. Existing frameworks were either qualitative (e.g., providing best practice checklists) or borrowed quantitative metrics from artificial intelligence (AI) fairness paradigms that strive for equality of AI performance, but do not prioritise performance for groups experiencing worse health outcomes. None of the existing frameworks provided a quantitative approach towards health equity assessment of machine learning performance that incorporates existing disparities in health outcomes.

Added value of this study

This work describes the framework for Health Equity Assessment of machine Learning performance (HEAL) designed to quantitatively assess the performance equity of health AI technologies via a four-step interdisciplinary process to understand and quantify domain-specific criteria, and the resulting HEAL metric. As an illustrative case study, we applied the HEAL framework to a dermatology AI model using a retrospective set of 5420 teledermatology cases.

Implications of all the available evidence

Analysis using the proposed HEAL framework showed that the dermatology AI model prioritised performance for race/ethnicity, sex (all conditions) and age (cancer conditions) subpopulations with respect to pre-existing health disparities. More work is needed to investigate ways of promoting equitable AI performance across age for non-cancer conditions and to better understand how AI models can contribute towards improving equity in health outcomes.

Introduction

Health equity is a major societal concern worldwide where health disparities are large, persistent and widening.[1–6] The causes of inequities in healthcare are multifactorial and precipitated by barriers at all levels of society. These include, but are not limited to, limitations in access to healthcare, differential clinical treatment, and even differences in diagnostic efficacy.[7,8] Recent advancement in artificial intelligence (AI) has accelerated the transition from academic research to bedside implementation of these technologies.[9] While there is optimism about the utility of AI technologies, introducing AI into clinical decision making carries the risk of exacerbating pre-existing inequities.[10] There is a growing sense of urgency within the academic, clinical, and regulatory communities to understand, monitor, and improve the effect of AI technologies through a health equity lens.[11–17]

Consequently, it is imperative that AI model development incorporates health equity considerations. Health equity is defined by public health organizations as everyone having a fair opportunity to be as healthy as possible (referred to as the 'health equity principle' below, see Table S1 for detailed definitions). Importantly, equity is not the same as equality.[18] Striving for health equity means we need to consider that individuals with larger barriers to improving their health require *more* and/or *different*, rather than equal, effort in order to experience this fair opportunity. Similarly, equity is not fairness as defined in the AI for healthcare literature. Whereas AI fairness often strives for equal performance of the AI technology across different patient populations,[19] this does not centre around the goal of prioritizing performance with respect to pre-existing health disparities.

We propose a methodology to assess whether health AI technologies prioritise performance for patient populations experiencing worse outcomes (Fig. 1A) that is complementary to existing fairness metrics. Specifically, we develop the Health Equity Assessment for machine Learning performance (HEAL) framework to quantitatively assess whether an AI tool's performance is equitable, defined as performing better for groups with worse average health outcomes as compared to others; anchoring on the principle that health equity should prioritise and measure model performance with respect to disparate health outcomes, which may be due to a number of factors that include structural inequities

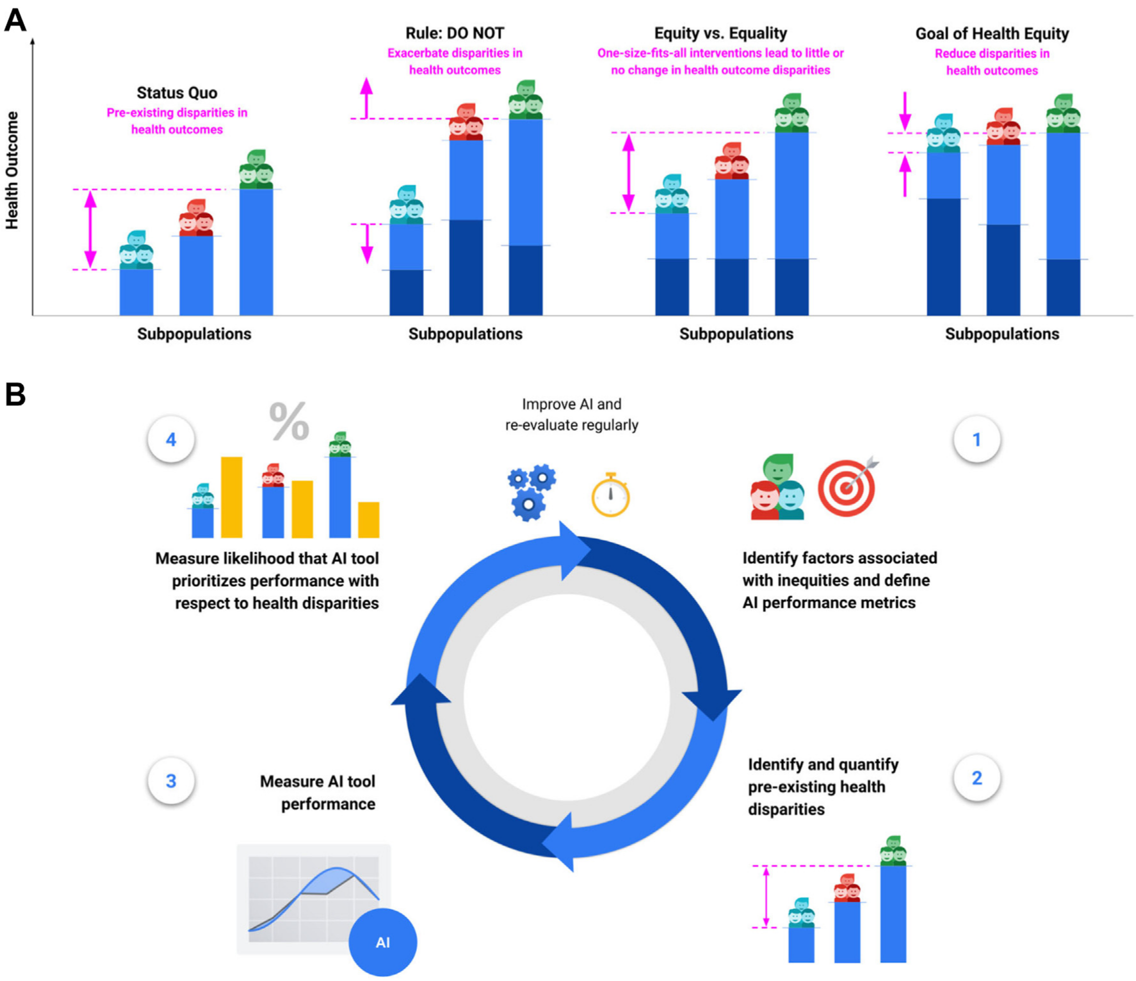

A
Health Outcome
Status Quo
Pre-existing disparities in health outcomes
Rule: DO NOT
Exacerbate disparities in health outcomes
Equity vs. Equality
One-size-fits-all interventions lead to little or no change in health outcome disparities
Goal of Health Equity
Reduce disparities in health outcomes
Subpopulations
Subpopulations
Subpopulations
Subpopulations
B
Improve AI and re-evaluate regularly
4
%
Measure likelihood that AI tool prioritizes performance with respect to health disparities
1
Identify factors associated with inequities and define AI performance metrics
3
Measure AI tool performance
AI
2
Identify and quantify pre-existing health disparities

Fig. 1: **Framework for Health Equity Assessment of machine Learning performance (HEAL). A**: An intervention promotes health equity if it contributes to reducing existing disparities in health outcomes. The light blue bars illustrate pre-existing health outcomes. The dark blue bars illustrate the effect of an intervention on pre-existing health outcomes. **B:** The process of estimating the likelihood that a health AI technology performs equitably entails four steps: (1) Identifying factors associated with health inequities and defining AI performance metrics; (2) Identifying and quantifying pre-existing health outcome disparities; (3) Measuring the performance of the AI tool for each subpopulation; (4) Measuring the likelihood that the novel AI tool prioritises performance with respect to health disparities. This 4-step process is designed to inform improvements for making the AI performance more equitable. HEAL metrics should be re-evaluated on a regular basis. Note that this is an iterative process. For example, the availability of health outcomes data in step (2) can inform the choice of demographic factors and brackets in step (1).

(e.g., demographic, social, cultural, political, economic, environmental and geographic). As an illustrative case study, we apply the framework to a dermatology AI model.[20]

With this work, we take a step towards encouraging explicit assessment of the health equity considerations of AI technologies, and encourage prioritization of efforts during AI development to reduce health inequities for subpopulations exposed to structural inequities that can precipitate disparate outcomes. While the present framework does not model causal relationships and, therefore, cannot quantify the actual impact a new AI technology will have on reducing health outcome disparities, the HEAL metric may help identify cases where model performance is not prioritised with respect to pre-existing health disparities.

Methods

Development of the HEAL framework

An interdisciplinary working group, including health equity researchers, social scientists, clinicians, bioethicists, statisticians, and AI researchers, developed the HEAL framework. The development process was iterative. It involved reviewing existing literature, ideating ways to address gaps in the literature, and grounding the feasibility of potential approaches in the context of a case study. The literature review suggested that existing AI fairness metrics typically do not account for pre-existing inequities,[21–25] and that existing health equity frameworks are either qualitative (e.g., providing best practice checklists)[14,26–29] or borrow quantitative metrics from AI fairness paradigms that strive for equality of AI performance, but do not prioritise performance for groups experiencing worse health outcomes.[30–32] The HEAL framework strives to bridge this gap by providing a four-step process (Fig. 1B) that AI model developers and researchers should undertake to produce a quantitative metric that assesses prioritization of model performance with respect to pre-existing health disparities.

Identify factors associated with health inequities and define metrics to quantify tool performance

This step may involve a combination of reviewing scientific literature and/or participatory methods, i.e., stakeholder engagement of clinicians, people with lived experience of structural inequities or the health condition being examined, to identify societal factors (e.g., demographic, social, cultural, political, economic, environmental and geographic) that have been associated with inequities within the given medical domain.[33] Each factor of inequity identified in this step is used as input to step (2) of the framework. Efforts should be taken to map the dynamic and complex relationship between the societal factors influencing health inequities.[34] Note that factors of inequity may vary by medical domain and the conditions of interest, so this step needs to be repeated when the HEAL framework is applied to a novel context.

Tool performance metrics should be chosen according to the task for which the AI tool is designed to assist (e.g., accuracy, sensitivity, specificity). Note that AI tool performance should be placed within the context of a care journey as part of mapping societal context. The AI tool's interdependence with other factors (e.g., condition incidence and discovery, information seeking, access to quality care, experiences in the clinical system) should be taken into account where possible.

Identify and quantify pre-existing health disparities

This step involves reviewing scientific literature and databases to identify existing disparities in health outcomes across factors of inequity, specifying the subpopulations associated with these disparities, and ranking subpopulations based on a quantitative health outcome measure. Note that steps (1) and (2) can be iterative. For example, the availability of health outcomes data in step (2) can inform the choice of factors and brackets in step (1). We use the term subpopulations because factors of inequity may extend beyond demographic factors. For example, the availability of outcome data and historical structural inequities may have strong associations with factors such as zip codes or insurance status in some domains.

Measure performance of AI tool

This step involves measuring the AI tool performance for each subpopulation. This can be done using a retrospective dataset or in a prospective manner. It requires a trusted reference standard and case-level metadata allowing sub-group analysis along the factors of inequity identified in the previous steps. AI tool performance per subpopulation is used as input to step (4) below.

Measure likelihood that AI tool prioritises performance with respect to health disparities

This step involves estimating the likelihood that the AI tool performs better for groups with poorer health outcomes using the HEAL metric. We next describe a case study in applying the HEAL framework.

Dermatology AI case study

The AI tool considered is a dermatology AI model for predicting potential dermatologic conditions based on photos of a skin concern and patient metadata. Dermatology is a suitable domain for a case study of the HEAL framework given the existing body of literature on factors associated with health inequities in dermatological care, and the risk that AI technologies may exacerbate those inequities.[35–38] The input to the model consists of three photos of a skin concern along with demographic information and a brief structured medical history, and the output consists of a ranked list of possible matching skin conditions. The AI model utilises a convolutional neural network similar to that described in prior work,[20] and was trained to classify 288 skin conditions using a development dataset of 29,087 cases (Table S2). Our study was conducted from October 2022 to January 2023.

Step 1: identifying factors associated with health inequities and defining tool performance metrics

After taking into consideration data availability and reviewing scientific literature to identify factors that have been associated with health inequities in dermatological care, we selected the following demographic factors: age, sex, race/ethnicity and Fitzpatrick skin type (FST).[35–38] FST is a classification system for human skin based on its response to ultraviolet (UV) radiation, particularly sunburn and tanning. The scale ranges from FST I to FST VI, with each type representing a different level of melanin production in the skin, eye, and hair, and sensitivity to UV light. Other factors relevant to dermatologic health outcomes (e.g., indicators of socioeconomic status) were not available in the datasets. Note that, in this study, we approximate FST as a dermatologist-provided *estimate* based on retrospective photo(s) of a skin concern (eFST; see Dataset section).

While this case study emphasises scientific literature review, we also encourage the use of participatory methods to identify and further elucidate potential factors of inequity, e.g., through focus groups involving people with lived experience.

We selected “top-3 agreement” as an appropriate performance metric to evaluate the dermatology AI. Top-3 agreement was used to evaluate AI performance because the AI output is a ranked list of possible matching medical conditions,[20] and is defined as the proportion of cases where at least one of the top-3 conditions suggested by the AI matches the reference diagnosis from a dermatologist panel. Although recommended as part of the process, we did not map the societal context influencing health inequities as they relate to the role of the AI tool within the care journey. The goal of this mapping process would be to understand how AI integration may impact existing inequities or beget new inequities in the clinical domain of interest. This is an important aspect of the framework that we hope to incorporate in the future.

Step 2: identifying and quantifying pre-existing health disparities

To determine a ranking of dermatologic health outcomes per demographic factor, we consulted the Global Burden of Disease (GBD) Results Tool,[39] a widely-used resource endorsed by the World Health Organization (WHO). We used GBD statistics for all skin conditions from the USA. Where available, we tabulated disability-adjusted life years (DALYs)[40] and determined the DALY rate (per 100,000) for each population. DALYs, defined as the number of years of healthy life lost, are the sum of years of life lost (YLLs) and years lived with disability (YLDs). For race/ethnicity, DALYs were not available from GBD data. Instead, we used the YLL rate as a health outcome measure.[41]

Skin cancers, also referred to as malignant neoplasms, are an area of particular concern in dermatology. We performed a sub-analysis on cancers to understand the impact of AI performance in these high-risk conditions. We used GBD categories “Non-melanoma skin cancer” and “Malignant skin melanoma” to estimate health outcomes for all cancers, and GBD category “Skin and subcutaneous diseases” for all non-cancer conditions. Nuances of these categories are described in Supplementary Methods section S4 “Details of GBD taxonomy”.

We also propose and tested a second approach to estimate health outcome rankings in the absence of available data on health outcome metrics to illustrate how health outcome measures can be derived from public data sets in geographic regions or medical domains where trusted health outcomes data is not readily available (Supplementary Methods section S2 “Care journey approach”). Conclusive health outcomes rankings could not be derived for FST due to the lack of publicly available data.

Step 3: measuring performance of AI tool

To understand the performance of the dermatology AI, we measured top-3 agreement by comparing AI’s predicted ranked conditions with the reference diagnosis on an evaluation dataset (see Dataset section for details), stratified by subpopulations based on age, sex, race/ethnicity and eFST. The reference diagnosis was established by aggregating the differential diagnoses from a panel of three US-board certified dermatologists per case via a voting procedure.[20] We report 95% confidence intervals computed using the Normal approximation of the binomial proportion confidence interval.

Step 4: measuring likelihood that AI tool prioritises performance with respect to health disparities

To quantify the HEAL metric for a health AI tool, our framework requires two inputs for each subpopulation: (1) a quantitative measure of pre-existing health outcomes, and (2) the AI performance. We first compute the anticorrelation between health outcomes and AI performances among all subpopulations for a given factor of inequity (e.g., race/ethnicity) as follows:

$$R = -\mathrm{corr}[\ (HO_{1,}\ HO_{2,\ \ldots,}\ HO_N),\ (AI_{1,}\ AI_{2,\ \ldots,}\ AI_N)\ ]$$

where **corr:** Spearman’s rank correlation.

HO_i: pre-existing health outcome for subpopulation i.

AI_i: AI tool performance for subpopulation i.

N: number of subpopulations considered.

Note that we define R to be the *negated* correlation coefficient, such that higher positive values of R correspond to a greater priority for AI performance for subpopulations with worse health outcomes: values of R close to 1.0 imply that AI performances are strongly anti-correlated with health outcomes. In other words, the subpopulation with the worst outcome has the highest AI performance, the subpopulation with the second-worst outcome achieves the second-best performance, and so forth. As illustrated in Fig. 1A, such a trend suggests that the AI tool prioritises model performance with respect to pre-existing health outcomes. Negative values close to −1.0 imply the opposite. Small values suggest that performance is mostly uncorrelated with health outcomes.

Next, the final HEAL metric, defined as p (R >0), measures how *likely* the AI technology is to prioritise performance with respect to pre-existing health outcomes. To obtain this metric, distributions of R were estimated via 9999 bootstrap samples. For each bootstrap sample, the entire set of patient cases was resampled with replacement, AI performance was calculated per subpopulation, and R was computed using the resulting AI performances in conjunction with health outcome measures. From this distribution we can empirically compute the HEAL metric, p (R >0), by counting the number of bootstrap samples with positive R. The HEAL metric ranges from 0 to 100%; a HEAL metric exceeding 50% implies that most of the bootstrap samples have R >0, suggesting higher likelihood of equitable performance. A HEAL metric less than 50% implies that most of the bootstrap samples have R ≤0, suggesting lower likelihood of equitable performance. Note that 50% is used as an illustrative value in this context to explain the concrete relationship between the HEAL metric and the underlying bootstrap sample, rather than serving as a binary cut-off threshold. Details can be found in Supplementary Methods section S1 "Health Equity Metric Considerations".

Statistical analysis

We used a multivariable logistic regression analysis to understand the effect of demographic variables (specifically age, sex, race/ethnicity, eFST) and dermatologic conditions on AI performance. Demographic variables and dermatologic conditions were used as independent variables and the correctness of the AI prediction was used as the binary dependent variable. The log odds for each variable were calculated along with the confidence intervals. Reference categories for the logistic regression analysis were 70+ years for age, male for sex, White for race/ethnicity, N/A for eFST, and "Other" for skin condition. Note that the multivariable logistic regression analysis is not an integral step of the HEAL framework, but rather was used in our case study to determine which subgroup analyses would be helpful to further elucidate via HEAL metric calculation. Rationales for the choices of reference groups in logistic regression analysis are orthogonal to specifics of the HEAL framework and are provided in Supplementary Methods section S3 "Reference categories for logistic regression analysis." Subgroup analyses were performed to determine HEAL metrics across age groups for cancer and non-cancer conditions separately and AI tool performance across FST subpopulations (Table S3). A supplementary intersectional analysis was performed to determine the HEAL metric across sex, age and race/ethnicity (Table S4). A sensitivity analysis was performed to derive health outcome rankings via an alternative approach (Table S6). Python packages NumPy version 1.26.2 and Statsmodels version 0.12.2 were used for data analysis and statistical methods.

Characteristics	Num. Cases (%)
No. of cases	5420
No. of images included in study	14,303
No. of patients included in study	4180
Female (%)	2980 (55.0%)
Age, median (25th, 75th percentiles)	55 (39, 67)
Race and ethnicity[a]	
White, non-Hispanic (%)	698 (32.2%)
Asian and Pacific Islander, non-Hispanic (%)	261 (12.0%)
Hispanic (%)	867 (40.0%)
Black, non-Hispanic (%)	291 (13.4%)
Other/Not specified (%)	51 (2.4%)
Estimated Fitzpatrick skin types	
I or II (%)	2621 (48.4%)
III or IV (%)	2026 (37.4%)
V or VI (%)	262 (4.8%)
N/A (%)	511 (9.4%)

Skin conditions based on reference diagnosis	Num. Cases (%)	Skin conditions based on reference diagnosis	Num. Cases (%)
Acne	135 (2.5%)	Psoriasis	93 (1.7%)
Actinic keratosis	357 (6.6%)	SCC/SCCIS[b]	493 (9.1%)
Allergic contact dermatitis	57 (1.1%)	SK/ISK	377 (7.0%)
Alopecia areata	46 (0.8%)	Scar condition	61 (1.1%)
Androgenetic alopecia	45 (0.8%)	Seborrheic dermatitis	46 (0.8%)
Basal cell carcinoma[b]	518 (9.6%)	Skin tag	37 (0.7%)
Cyst	111 (2.0%)	Stasis dermatitis	36 (0.7%)
Eczema	159 (2.9%)	Tinea	35 (0.6%)
Folliculitis	53 (1.0%)	Tinea versicolor	31 (0.6%)
Hidradenitis	44 (0.8%)	Urticaria	17 (0.3%)
Lentigo	152 (2.8%)	Verruca vulgaris	44 (0.8%)
Melanocytic nevus	563 (10.4%)	Vitiligo	41 (0.8%)
Melanoma[b]	195 (3.6%)	Other condition	1636 (30.2%)
Post-inflammatory hyperpigmentation	38 (0.7%)		

Enrichment was performed to mitigate skew towards common demographics and dermatologic conditions, and additionally to include all available cases with race/ethnicity information. The table includes 26 common skin conditions, representing 80% of cases seen in primary care,[20] and an additional category grouping other conditions. [a]Race/ethnicity information was unavailable for the skin cancer dataset. Case counts and percentages for race/ethnicity therefore reflect only cases from the tele-dermatology dataset where race/ethnicity was generally available. [b]Skin cancers.

***Table 1:* Patient characteristics in the curated evaluation dataset.**

Dataset

To measure AI performance, we curated an evaluation dataset of 5420 store-and-forward cases (Table 1) with reference diagnoses, which was sampled from two sources. The first was a tele-dermatology dataset from the USA (California, Hawaii) of mostly low-to-medium risk conditions, and with age, sex and self-reported race/ethnicity information available. The second dataset was from several skin cancer clinic sites in Australia to enrich for malignant neoplasms, with age and sex, but no race/ethnicity information. Patients with race/ethnicity information were of a single race/ethnicity group, i.e., either of Hispanic ethnicity or belonging to one of four racial groups (Black, Asian and Pacific Islander, American Indian and Alaska Native, or White). No cases of mixed race/ethnicity were available for evaluation. eFST was estimated for each case via the majority vote among three independent dermatologist assessments based on retrospective photos. All cases came from 2020 or before, with exact dates not available due to de-identification.

Race/Ethnicity	N	Health outcomes		AI tool performance		HEAL metric
		YLLs (per 100,000)	Ranking	Top-3 agreement (95% CI)	Ranking	
White, non-Hispanic	698	223.7	4	80.7% (77.5, 83.5)	2	R = 0.4 p (R >0) = 80.5%
Asian and Pacific Islander, non-Hispanic	261	45.2	1	76.6% (71.0, 81.6)	4	
Hispanic	867	70.4	2	84.7% (82.1, 87.0)	1	
Black, non-Hispanic	291	131.4	3	80.1% (75.0, 84.5)	3	
Other/Not specified	51	N/A	N/A	80.4% (66.9, 90.2)	N/A	
R/E info unavailable for dataset	3252	N/A	N/A	72.0% (70.4, 73.5)	N/A	
Age group (yrs)	**N**	**Health outcomes**		**AI tool performance**		**HEAL metric**
		DALYs (per 100,000)	**Ranking**	**Top-3 agreement (95% CI)**	**Ranking**	
20–24	309	557.4	1	89.0% (85.0, 92.3)	1	R = −1.0 p (R >0) = 0.0%
25–49	1941	606.1	2	78.8% (76.9, 80.6)	2	
50–69	2052	942.6	3	73.5% (71.5, 75.4)	3	
70+	1118	1418.7	4	71.6% (68.8, 74.2)	4	
Sex	**N**	**Health outcomes**		**AI tool performance**		**HEAL metric**
		DALYs (per 100,000)	Ranking	Top-3 agreement (95% CI)	Ranking	
Female	2980	832.6	2	76.6% (75.0, 78.1)	1	R = 1.0 p (R >0) = 92.1%
Male	2440	830.6	1	75.0% (73.2, 76.7)	2	

Table 2: **HEAL metrics for all dermatologic conditions including health outcomes (DALYs or YLLs per 100,000), AI performances (top-3 agreement), and rankings for health outcomes and tool performances, with breakdowns by race/ethnicity, age group, and sex.**

We applied exclusion criteria and sampling to enrich the dataset for demographics and conditions traditionally underrepresented in datasets to maximise sample sizes for each intersectional subpopulation. We excluded patients under 20 years of age to match age brackets from GBD data. For the tele-dermatology dataset, we included all cases to maximise the case count with race/ethnicity information available. For the skin cancer clinic dataset, we formed distinct intersections across 4 age groups (20–24 yrs, 25–49 yrs, 50–69 yrs, 70 + yrs), 2 sexes, 6 eFSTs, and 60 skin conditions, and randomly sampled up to 20 cases from each intersection. See Table S4 for the number of cases across each intersectional subpopulation across age (binarised), sex and race/ethnicity.

Table 1 summarises patient characteristics in the curated evaluation dataset. Details on case exclusions can be found in Figure S3, and a comprehensive list of all dermatologic conditions in the dataset is provided in Table S5.

Ethics

Given the retrospective nature of this study and the use of de-identified datasets, the need for further review was waived by the Advarra Institutional Review Board (IRB).

Role of the funding source

This study was funded by Google LLC and the majority of co-authors had a Google affiliation while contributing to this work, including study design, data collection, data analyses, interpretation, and writing of the report.

Results

Table 2 summarises health outcomes across age, sex and race/ethnicity groups for all dermatologic conditions. For race/ethnicity, we observed the best outcomes (lowest YLL rate) for the Asian and Pacific Islander subpopulation (45.2 YLLs), followed by Hispanic (70.4 YLLs), Black (131.4 YLLs) and White (223.7 YLLs). For sex, the GBD data suggest that the male subpopulation (830.6 DALYs) have slightly better health outcomes than the female subpopulation (832.6 DALYs). With respect to age, across all conditions, the oldest age group of 70+ year-olds experiences the worst health outcomes, followed by 50–69 year-olds. Across all conditions (Table 2) and for cancers (Table 3A), the trend continues with health outcomes improving for 25–49 year-olds and 20–24 year-olds. For non-cancer conditions (Table 3B), the ranking is reversed for the two youngest age groups.

In Table 2, we summarise the HEAL analysis across age, sex, and race/ethnicity. For race/ethnicity, the analysis was performed on four subpopulations (Hispanic, Black, Asian/Pacific Islander, White), and the HEAL metric was 80.5%. For sex, the HEAL metric was 92.1%.

For age, we observed a HEAL metric of 0.0% computed across four subpopulations (20–24 years, 25–49 years, 50–69 years, 70+ years), suggesting a low likelihood of prioritizing performance with respect to health disparities across age groups. Logistic regression analysis (Fig. 2) revealed that, in addition to age, certain dermatologic conditions had a significant effect on AI performance. For example, the AI performed more accurately for some cancers (basal cell carcinoma,

3. A Cancer conditions.

Age group (yrs)	N	Health outcomes		AI tool performance		HEAL metric
		DALYs (per 100,000)	Ranking	Top-3 agreement (95% CI)	Ranking	
25–49	178	67.4	1	73.0% (65.9, 79.4)	3	R = 1.0 p (R >0) = 73.8%
50–69	543	257.1	2	79.7% (76.1, 83.0)	2	
70+	495	475.7	3	83.4% (79.9, 86.6)	1	

3. B Non-cancer conditions

Age group (yrs)	N	Health outcomes		AI tool performance		HEAL metric
		DALYs (per 100,000)	Ranking	Top-3 agreement (95% CI)	Ranking	
20–24	308	545.3	2	89.0% (84.9, 92.2)	1	R = −0.8 p (R >0) = 0.0%
25–49	1763	538.6	1	79.4% (77.4, 81.2)	2	
50–69	1509	685.5	3	71.2% (68.9, 73.5)	3	
70+	623	943.0	4	62.1% (58.2, 65.9)	4	

Table 3: **HEAL metrics for cancer (3.A) and non-cancer (3.B) dermatologic conditions including health outcomes (DALYs per 100,000) and AI performances (top-3 agreement) across age groups. For cancer conditions, the youngest age group (20–24 yrs, 12.1 DALYs per 100,000) was excluded due to insufficient sample size (N = 1).**

squamous cell carcinoma) and less accurately for other conditions (e.g., cyst) compared to a reference category of other less common conditions, when controlling for demographic factors.

To further explore the relationship between dermatologic conditions and age groups, we performed HEAL analysis across age separately for cancer (Table 3A) and non-cancer conditions (Table 3B). For cancers, we observed a HEAL metric of 73.8%, suggesting a high likelihood of prioritizing performance with respect to health disparities. However, among non-cancer conditions, the HEAL metric was 0.0%. The group of 70+ year-olds had the poorest health outcomes paired with lowest AI performance. Through this analysis, we were able to identify a specific disease group (non-cancers) and factor (age) across which AI performance needs to be improved to increase the likelihood of equitable performance.

In the absence of quantitative health outcomes data by eFST subgroups, we report AI performance by eFST in Table S3, suggesting a trend towards greater AI performance for people with darker skin.

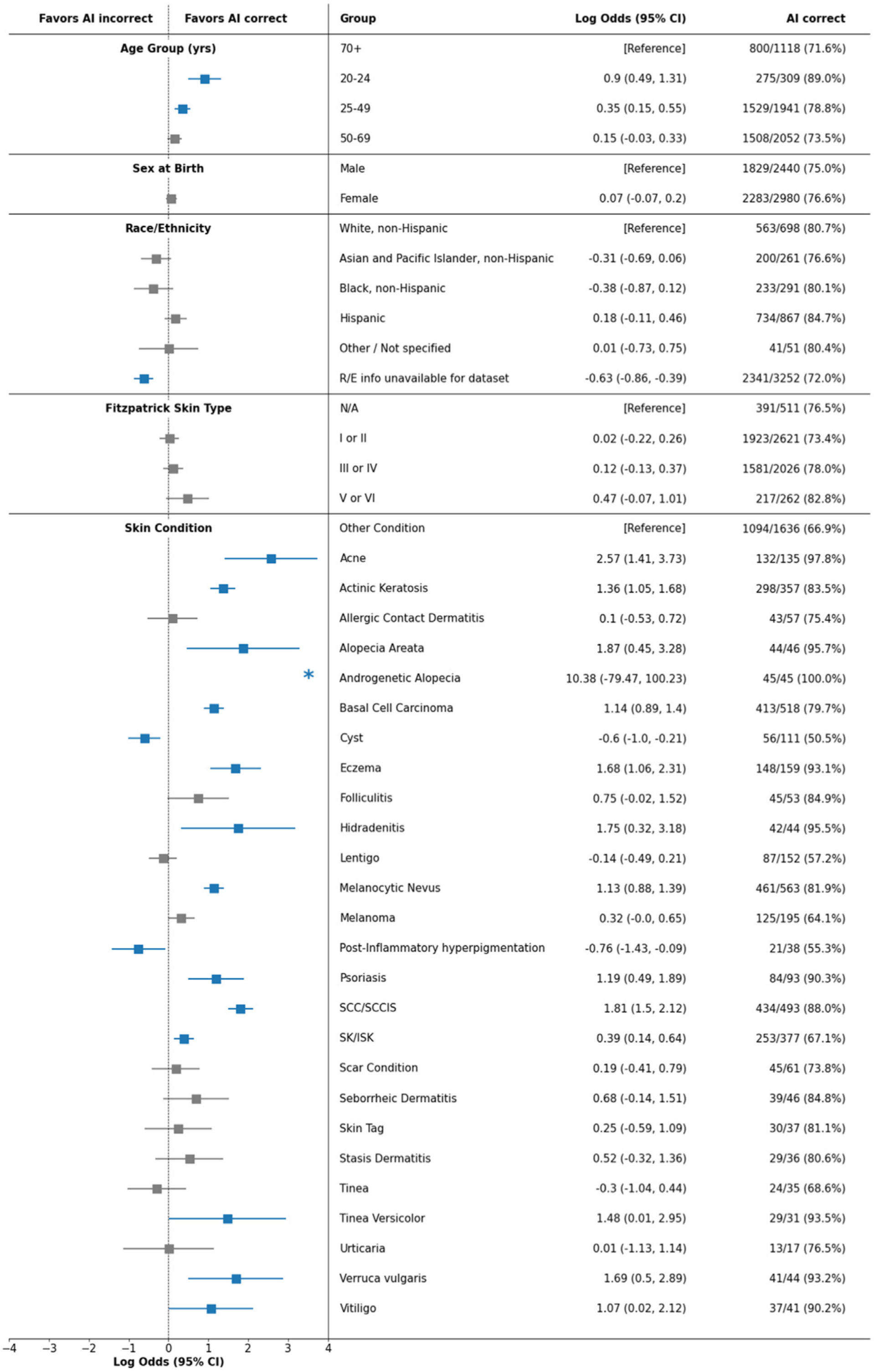
We recognise the importance of intersectionality in health equity. In Table S4, we report an extended HEAL analysis across intersections of age, sex and race/ethnicity enabled by the fine-grained GBD health outcomes measures (YLL rate). Overall, we observed a HEAL metric of 17.0%. To dive deeper into the results, we focused on intersections ranked in the lower half for both health outcomes and AI performance and identified subpopulations for which AI performance needs to be improved to increase the likelihood of prioritizing performance with respect to health disparities: female/50+/Hispanic, female/50+/Black, female/50+/White, male/20–49/White, and male/50+/Asian and Pacific Islander.

Discussion

Here, we proposed the HEAL framework to quantify the likelihood of equitable performance for AI tools in health. The framework comes with a metric grounded in prioritizing performance with respect to pre-existing health disparities, that is straightforward to implement and interpret. The HEAL case study suggests that the dermatology AI prioritises performance with respect to pre-existing health disparities across race/ethnicity and sex and, for cancers, across age groups. Finally, we identified subpopulations where further AI improvements are needed: across age groups for non-cancers.

Other work has emphasised the need for operationalizing equity and fairness in AI for healthcare.[11,12] Recent work addressing this call has discussed ethical considerations of fairness and equity in the context of AI for healthcare,[42–47] suggested best practices to incorporate health equity in the algorithm development lifecycle,[14,26–29,48] and proposed operational definitions.[30–32] Existing operational definitions have largely borrowed from the AI fairness literature,[21–24] proposing metrics based on statistical parity in AI performance across subpopulations. Our contribution to this space is to anchor quantitative evaluation of AI on pre-existing health disparities. The framework can potentially be extended beyond healthcare and beyond the use of AI-driven tools given reliable ways to measure outcomes.

The framework can be applied to a range of medical domains, AI tasks and personal or contextual attributes. Some steps need to be adapted when applied in a new context. For example, demographic factors and health outcomes rankings can be reused when applying the framework to a different AI model within the same domain. The choice of appropriate performance metrics depends on the AI task. We use top-3 agreement as an



	Group	Log Odds (95% CI)	AI correct
Age Group (yrs)	70+	[Reference]	800/1118 (71.6%)
	20-24	0.9 (0.49, 1.31)	275/309 (89.0%)
	25-49	0.35 (0.15, 0.55)	1529/1941 (78.8%)
	50-69	0.15 (-0.03, 0.33)	1508/2052 (73.5%)
Sex at Birth	Male	[Reference]	1829/2440 (75.0%)
	Female	0.07 (-0.07, 0.2)	2283/2980 (76.6%)
Race/Ethnicity	White, non-Hispanic	[Reference]	563/698 (80.7%)
	Asian and Pacific Islander, non-Hispanic	-0.31 (-0.69, 0.06)	200/261 (76.6%)
	Black, non-Hispanic	-0.38 (-0.87, 0.12)	233/291 (80.1%)
	Hispanic	0.18 (-0.11, 0.46)	734/867 (84.7%)
	Other / Not specified	0.01 (-0.73, 0.75)	41/51 (80.4%)
	R/E info unavailable for dataset	-0.63 (-0.86, -0.39)	2341/3252 (72.0%)
Fitzpatrick Skin Type	N/A	[Reference]	391/511 (76.5%)
	I or II	0.02 (-0.22, 0.26)	1923/2621 (73.4%)
	III or IV	0.12 (-0.13, 0.37)	1581/2026 (78.0%)
	V or VI	0.47 (-0.07, 1.01)	217/262 (82.8%)
Skin Condition	Other Condition	[Reference]	1094/1636 (66.9%)
	Acne	2.57 (1.41, 3.73)	132/135 (97.8%)
	Actinic Keratosis	1.36 (1.05, 1.68)	298/357 (83.5%)
	Allergic Contact Dermatitis	0.1 (-0.53, 0.72)	43/57 (75.4%)
	Alopecia Areata	1.87 (0.45, 3.28)	44/46 (95.7%)
	Androgenetic Alopecia	10.38 (-79.47, 100.23)	45/45 (100.0%)
	Basal Cell Carcinoma	1.14 (0.89, 1.4)	413/518 (79.7%)
	Cyst	-0.6 (-1.0, -0.21)	56/111 (50.5%)
	Eczema	1.68 (1.06, 2.31)	148/159 (93.1%)
	Folliculitis	0.75 (-0.02, 1.52)	45/53 (84.9%)
	Hidradenitis	1.75 (0.32, 3.18)	42/44 (95.5%)
	Lentigo	-0.14 (-0.49, 0.21)	87/152 (57.2%)
	Melanocytic Nevus	1.13 (0.88, 1.39)	461/563 (81.9%)
	Melanoma	0.32 (-0.0, 0.65)	125/195 (64.1%)
	Post-Inflammatory hyperpigmentation	-0.76 (-1.43, -0.09)	21/38 (55.3%)
	Psoriasis	1.19 (0.49, 1.89)	84/93 (90.3%)
	SCC/SCCIS	1.81 (1.5, 2.12)	434/493 (88.0%)
	SK/ISK	0.39 (0.14, 0.64)	253/377 (67.1%)
	Scar Condition	0.19 (-0.41, 0.79)	45/61 (73.8%)
	Seborrheic Dermatitis	0.68 (-0.14, 1.51)	39/46 (84.8%)
	Skin Tag	0.25 (-0.59, 1.09)	30/37 (81.1%)
	Stasis Dermatitis	0.52 (-0.32, 1.36)	29/36 (80.6%)
	Tinea	-0.3 (-1.04, 0.44)	24/35 (68.6%)
	Tinea Versicolor	1.48 (0.01, 2.95)	29/31 (93.5%)
	Urticaria	0.01 (-1.13, 1.14)	13/17 (76.5%)
	Verruca vulgaris	1.69 (0.5, 2.89)	41/44 (93.2%)
	Vitiligo	1.07 (0.02, 2.12)	37/41 (90.2%)

Fig. 2: Logistic regression analysis for understanding the effect of demographics factors and skin conditions on the correctness of AI predictions. Blue indicates statistical significance. Asterisk (*) indicates that there was no separation of cases into distinct outcomes within a given sub-population and, as a result, log odds have extreme values. Fitzpatrick skin type was estimated for each case via the majority vote among three independent dermatologist assessments based on retrospective photos.

example of a metric appropriate for evaluating ranked lists of predictions; for binary classification tasks, other metrics, such as the area under the receiver-operating characteristic (ROC) curve (AUC), may be more suitable. While our case study focuses on specific demographics, the framework can be applied to any grouping dimension (e.g., social determinants of health) along which health outcomes and AI performance can be quantified. Note that application of the framework in real-world monitoring settings, will require nuanced consideration of additional aspects such as comorbidity.

For holistic evaluation, the HEAL metric should be contextualised alongside competing performance factors (e.g., computational efficiency and data privacy), ethical values (e.g., doing no harm and increasing overall utility), and forms of bias that may influence the results (e.g., selection bias or differences in representativeness of the evaluation data across demographic groups).[49] In other words, if the HEAL metric is to be employed, it cannot be interpreted and acted upon in isolation, but needs to be combined with other metrics and considerations based on the target intended application and population. For example, the HEAL metric can be artificially improved by deliberately reducing AI performance for the most advantaged subpopulation until AI performance for that subpopulation is worse than all other subpopulations. For illustrative purposes, given subpopulations A and B where A has worse health outcomes than B, consider the choice between two models: Model 1 (M1) performs 5% better for subpopulation A than for subpopulation B. Model 2 (M2) performs 5% worse on subpopulation A than B. The HEAL metric would be higher for M1. However, M1 has absolute performances of just 75% vs 70% for subpopulations A and B respectively, while M2 has absolute performances of 75% and 80% for subpopulations A and B respectively. Choosing M1 over M2 would lead to worse overall performance for all subpopulations. This is ethically problematic because some subpopulations are rendered worse-off while no subpopulation is better-off. Accordingly, the HEAL metric ought to be used alongside a Pareto condition that restricts model improvements to just those improvements such that each subpopulation is at least as well-off and some subpopulation is better-off compared to the status quo. This is to avoid situations where applying the framework leads to poorer AI performance for every subpopulation. To mitigate settings where application of the HEAL framework leads to violation of this Pareto condition, it also may be reasonable to consider an adaptation of the HEAL metric that correlates pre-existing health outcomes with performance *improvements*, rather than with the AI's absolute performance values. This ensures safe performance levels for all subpopulations while focusing further performance improvement efforts specifically on subpopulations exposed to structural inequities that can precipitate worse health outcomes. While the detailed operationalization of the Pareto principle for the HEAL metric is beyond the scope of this work, possible techniques for specifically improving AI model performance in subpopulations of interest may include targeted data collection or modified training procedures (such as data sampling or loss function weighting). Future research may explore and recommend concrete ways to contextualise the HEAL metric alongside competing constraints in the process of model development and deployment decisions.

The HEAL framework, in its current form, assesses the likelihood that an AI model prioritises performance for subpopulations with respect to pre-existing health disparities for specific subpopulations exposed to structural inequities that can precipitate disparate health outcomes, which differs from the goal of understanding whether AI can help reduce disparities in outcomes across subpopulations (since the latter requires a causal understanding of steps in the care journey that happen both before and after use of the AI model). Future refinements to the HEAL framework should work to address this gap. One potential approach to doing so involves developing a dynamic hypothesis, in the form of a system dynamics causal model, of the underlying socio-technical context and structure that produces specific disparate health outcomes and in which the AI tool would operate.[33,34,50] Such a model would incorporate AI tool outputs as key factors that can monotonically affect other contextual factors, such as trust and screening, and allow the evaluation of the impact of AI tool performance on downstream health outcomes. An important aspect of this approach is to consider the systemic and structural causes of health disparities and the potential for the AI tool to counteract them. We further acknowledge that application of the HEAL framework requires a grounding in existing literature which, in itself, may encode structural inequities, and as such may impose limitations on how accurately historic disparities can be captured. That said, we encourage careful review of the resources used to derive health outcome measures to avoid perpetuation of structural inequities as may be encoded in the literature.

This study has limitations. First, even though the dataset was sampled to balance demographics and dermatologic conditions, there were still fewer cases of American Indian/Alaska Native populations and eFST V and VI. For breakdowns by race/ethnicity, the available GBD health outcomes were limited to YLLs (rather than DALYs), whereas the subset of the evaluation dataset sourced from skin cancer clinics did not have race/ethnicity information available. Second, while we consider DALYs a useful health outcome measure, we recognise that limitations of the measure pertaining to disability weightings have been critiqued, and acknowledge the challenge in selecting appropriate health outcome measures when applying the HEAL framework, especially in geographic regions or medical domains where public health data sets are scarce.[51] We acknowledge that HEAL is sensitive with respect to the choice of a specific health outcome measure, and therefore strongly encourage careful consideration of which outcome measure to choose and cautious interpretation of HEAL metric values in that context. Additional applications of this framework could be considered that use information such as structural disadvantage instead of health outcomes to determine the priority order. Third, although we were able to perform eFST annotation on the evaluation dataset, there is no publicly available health outcomes data across FSTs. While FST is a widely-accepted skin type classification system in dermatology, limitations such as inadequate representation of black and brown skin tones are known. Further, the eFST labels in this dataset did not contain self-reported tanning propensity, but were instead estimated retrospectively by dermatologists based on visible factors such as healthy skin surrounding the condition, hair color, and tan lines. Finally, patient cases and health outcomes representing intersectional race/ethnicity groups (e.g., Black people of Hispanic ethnicity) or finer granularity (e.g., subpopulations within the Asian category)[52] were not available for evaluation. Further study to address nuanced considerations of representativeness is warranted, including research into the effect that sampling techniques for the evaluation dataset as well as subpopulation-specific disease prevalence may have on calculated metric values.

To conclude, the HEAL framework enables a quantitative assessment of the likelihood that health AI technologies prioritise performance with respect to health disparities. The case study demonstrated how to apply the framework in the dermatological domain, highlighting high likelihood that model performance is prioritised with respect to health disparities across sex and race/ethnicity, but also the potential for improvements for non-cancer conditions across age. The case study also illustrated limitations in our ability to apply all recommended aspects of the framework (e.g., mapping societal context, availability of data), thus highlighting the complexity of health equity considerations of AI tools. This work is a proposal towards addressing a grand challenge for AI and health equity, and may provide a useful evaluation framework not only during model development, but during pre-implementation and real-world monitoring stages, e.g., in the form of health equity dashboards. We hold that the strength of the HEAL framework is in its future application to various AI tools and use cases and its refinement in the process. Finally, we acknowledge that a successful approach towards understanding the impact of AI technologies on health equity needs to be more than a set of metrics.[53,54] It will require a set of goals agreed upon by a community that represents those who will be most impacted by a model.

Contributors

MS, TS, MP, CS, YL, IH, PHCC, EW, and CM contributed to the conception and design of the work. MS, TS, PHCC and PS contributed to the acquisition of data. MS, TS, PHCC, QX, MP, and RJ contributed to the analysis of data. MS, TS, MP, SF, JL, CH, FT, and PHCC contributed to the interpretation of data. MS, TS, and PHCC contributed to the creation of new software used in the work. MS, TS, MP, HCL, GK, YL, SP, DM, PS, SV, YL, IH, and PHCC contributed to drafting and revising the manuscript. PB, LP, CM, YM, GC, DW, SV, CS, IH, and PHCC contributed to the strategic direction and funding of the work. MS, PS, CS, YL, IH, and PHCC contributed to paper organization and team logistics. MS and TS have accessed and verified the underlying data. All authors read and approved the final version of the manuscript.

Data sharing statement

The de-identified dermatology case data used in this study are not publicly available due to restrictions in the data-sharing agreements. The health outcomes data are from the Institute for Health Metrics and Evaluation (IHME). Used with permission. All rights reserved.

Declaration of interests

This study was funded by Google LLC. MS, TS, HC, EW, SP, DM, RJ, GK, YL, SF, QX, CH, PS, FT, PB, LHP, CHM, YM, GSC, DW, SV, CS, YL, IH, PHCC are current or former employees of Google and own stock as part of the standard compensation package. MP and JL are paid consultants of Google.

Acknowledgements

The authors would like to acknowledge Ting-An Lin, Aaron Loh, Morgan Du, Jenny Rizk, Renee Wong, Ashley Carrick, Preeti Singh, Annisah Um'rani, Jessica Schrouff, Alexander Brown, and Anna Iurchenko for their support of this project. We appreciate the input of Naama Hammel, Ebony Respress, Perry Payne, and Michael Howell for their feedback on the manuscript. Lastly, we would like to thank the Institute for Health Metrics and Evaluation for licensing the GBD dataset.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102479.

References

1 Bibbins-Domingo K. The urgency of now and the responsibility to do more-my commitment for JAMA and the JAMA network. *JAMA*. 2022;328(1):21–22. https://doi.org/10.1001/jama.2022.11108.

2 Chew M, Das P, Aujla M, Horton R. Advancing racial and ethnic equity in science, medicine, and health: a call for papers. *Lancet*. 2021;398(10308):1287–1289. https://doi.org/10.1016/S0140-6736(21)02095-X.

3 Fontanarosa PB, Flanagin A, Ayanian JZ, et al. Equity and the JAMA network. *JAMA*. 2021;326(7):618–620. https://doi.org/10.1001/jama.2021.9377.

4 Das P, Aujla M, GRacE. Racial and ethnic equality - time for concrete action. *Lancet*. 2020;396(10257):1055–1056. https://doi.org/10.1016/S0140-6736(20)32077-8.

5 Richards, Franco, Bloom, Others. A commitment to equality, diversity, and inclusion for BMJ and our journals. BMJ Blogs. https://blogs.bmj.com/bmj/2021/07/23/a-commitment-to-equality-diversity-and-inclusion-for-bmj-and-our-journals/; 2021.

6 Equity, diversity, and inclusion collection. Lancet. https://www.thelancet.com/equity-diversity-inclusion/collection. Accessed August 4, 2023.

7 Penman-Aguilar A, Talih M, Huang D, Moonesinghe R, Bouye K, Beckles G. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *J Public Health Manag Pract*. 2016;22(Suppl 1):S33–S42. https://doi.org/10.1097/PHH.0000000000000373.

8 Racial bias in pulse oximetry measurement. *N Engl J Med*. 2021;385(26):2496. https://doi.org/10.1056/NEJMx210003.

9 U.S. Food & Drug Administration. Artificial intelligence and machine learning (AI/ML)-Enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices; 2022. Accessed November 14, 2022.

10 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):447–453. https://doi.org/10.1126/science.aax2342.

11 Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform*. 2021;28(1). https://doi.org/10.1136/bmjhci-2020-100289.

12 Parbhoo S, Wawira Gichoya J, Celi LA, de la Hoz MÁA. Operationalising fairness in medical algorithms. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2022-100617.

13 Striving for health equity with machine learning. *Nat Mach Intell*. 2021;3(8):653. https://doi.org/10.1038/s42256-021-00385-0.

14 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):866–872. https://doi.org/10.7326/M18-1990.

15 *Artificial intelligence/machine learning (AI/ML)-Based software as a medical device (SaMD) action plan*. U.S. Food & Drug Administration (FDA); 2021. https://www.fda.gov/media/145022/download.

16 Hammond A, Jain B, Celi LA, Stanford FC. An extension to the FDA approval process is needed to achieve AI equity. *Nat Mach Intell*. 2023;5:96–97. https://doi.org/10.1038/s42256-023-00614-8.

17 The Lancet Digital Health. New resolutions for equity. *Lancet Digit Health*. 2022;4(1):e1. https://doi.org/10.1016/S2589-7500(21)00280-6.

18 Braveman P, Arkin E, Orleans T, Proctor D, Plough A. *What is health equity? And what difference does a definition make?* Robert Wood Johnson Foundation; 2017.

19 Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1–35. https://doi.org/10.1145/3457607.

20 Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. 2020;26(6):900–908. https://doi.org/10.1038/s41591-020-0842-3.

21 Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*. 2017;5(2):153–163. https://doi.org/10.1089/big.2016.0047.

22 Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. *arXiv [csLG]*; 2016. http://arxiv.org/abs/1609.05807.

23 Narayanan A. Translation tutorial: 21 fairness definitions and their politics. New York, USA. In: *Proc. Conf. Fairness Accountability Transp*. 1170. 2018:3.

24 Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the international Workshop on software fairness. *FairWare '18*. Association for Computing Machinery; 2018:1–7. https://doi.org/10.1145/3194770.3194776.

25 Fazelpour S, Lipton ZC, Danks D. Algorithmic fairness and the situated dynamics of justice. *Can J Philos*. 2022;52(1):44–60. https://doi.org/10.1017/can.2021.24.

26 Dankwa-Mullan I, Scheufele EL, Matheny ME, et al. A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. *J Health Care Poor Underserved*. 2021;32(2):300–317.

27 Sikstrom L, Maslej MM, Hui K, Findlay Z, Buchman DZ, Hill SL. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100459.

28 Rojas JC, Fahrenbach J, Makhni S, et al. Framework for integrating equity into machine learning models: a case study. *Chest*. 2022;161(6):1621–1627. https://doi.org/10.1016/j.chest.2022.02.001.

29 Cerrato P, Halamka J, Pencina M. A proposal for developing a platform that evaluates algorithmic equity and accuracy. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100423.

30 Zink A, Rose S. Identifying undercompensated groups defined by multiple attributes in risk adjustment. *BMJ Health Care Inform*. 2021;28(1). https://doi.org/10.1136/bmjhci-2021-100414.

31 Straw I, Wu H. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100457.

32 Foryciarz A, Pfohl SR, Patel B, Shah N. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100460.

33 Prabhakaran V, Martin D Jr. Participatory machine learning using community-based system dynamics. *Health Hum Rights*. 2020;22(2):71–74. https://www.ncbi.nlm.nih.gov/pubmed/33390696.

34 Kuhlberg J, Headen I, Ballard E, Martin D Jr. Advancing community engaged approaches to identifying structural drivers of racial bias in health diagnostic algorithms. In: *Conference record the 2020 conference of the system dynamics society*; 2020. https://d4bl.org/reports/89-advancing-community-engaged-approaches-to-identifying-structural-drivers-of-racial-bias-in-health-diagnostic-algorithms.

35 Brady J, Kashlan R, Ruterbusch J, Farshchian M, Moossavi M. Racial disparities in patients with melanoma: a multivariate survival analysis. *Clin Cosmet Investig Dermatol*. 2021;14:547–550. https://doi.org/10.2147/CCID.S311694.

36 Nelson B. How dermatology is failing melanoma patients with skin of color: unanswered questions on risk and eye-opening disparities in outcomes are weighing heavily on melanoma patients with darker skin. *Cancer Cytopathol*. 2020;128(1):7–8. https://doi.org/10.1002/cncy.22229.

37 Orenstein LAV, Nelson MM, Wolner Z, et al. Differences in outpatient dermatology encounter work relative value units and net payments by patient race, sex, and age. *JAMA Dermatol*. 2021;157(4):406–412. https://doi.org/10.1001/jamadermatol.2020.5823.

38 Tripathi R, Knusel KD, Ezaldein HH, Scott JF, Bordeaux JS. Association of demographic and socioeconomic characteristics with differences in use of outpatient dermatology services in the USA. *JAMA Dermatol*. 2018;154(11):1286–1291. https://doi.org/10.1001/jamadermatol.2018.3114.

39 GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2020;396(10258):1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.

40 Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020 ; summary*; 1996. https://books.google.com/books/about/The_Global_Burden_of_Disease.html?hl=&id=lzQfAQAAIAAJ.

41 GBD US Health Disparities Collaborators. Cause-specific mortality by county, race, and ethnicity in the USA, 2000-19: a systematic analysis of health disparities. *Lancet*. 2023;402(10407):1065–1082. https://doi.org/10.1016/S0140-6736(23)01088-7.

42 Bærøe K, Gundersen T, Henden E, Rommetveit K. Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health Care Inform*. 2022;29(1). https://doi.org/10.1136/bmjhci-2021-100445.

43 Starke G, De Clercq E, Elger BS. Towards a pragmatist dealing with algorithmic bias in medical machine learning. *Med Health Care Philos*. 2021;24(3):341–349. https://doi.org/10.1007/s11019-021-10008-5.

44 Grote T, Keeling G. On algorithmic fairness in medical practice. *Camb Q Healthc Ethics*. 2022;31(1):83–94. https://doi.org/10.1017/S0963180121000839.

45 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics*. 2020;46(3):205–211. https://doi.org/10.1136/medethics-2019-105586.

46 Grote T, Keeling G. Enabling fairness in healthcare through machine learning. *Ethics Inf Technol*. 2022;24(3):39. https://doi.org/10.1007/s10676-022-09658-7.

47 Lin TA, Chen PHC. Artificial intelligence in a structurally unjust society. *FPQ*. 2022;8(3/4). https://ojs.lib.uwo.ca/index.php/fpq/article/view/14191. Accessed January 30, 2023.

48 Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. *NPJ Digit Med*. 2022;5(1):1–6. https://doi.org/10.1038/s41746-022-00663-0.

49 Petersen E, Holm S, Ganz M, Feragen A. The path toward equal performance in medical machine learning. *Patterns (N Y)*. 2023;4(7):100790. https://doi.org/10.1016/j.patter.2023.100790.

50 Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health*. 2022;4(5):e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6.

51 Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years. *Lepr Rev*. 2000;71(2):123–127. https://doi.org/10.5935/0305-7518.20000015.

52 Gordon NP, Lin TY, Rau J, Lo JC. Aggregation of Asian-American subgroups masks meaningful differences in health and health risks among Asian ethnicities: an electronic health record based cohort study. *BMC Public Health*. 2019;19(1):1551. https://doi.org/10.1186/s12889-019-7683-3.

53 Mbakwe AB, Lourentzou I, Celi LA, Wu JT. Fairness metrics for health AI: we have a long way to go. *eBioMedicine*. 2023;90:104525. https://doi.org/10.1016/j.ebiom.2023.104525.

54 Gallifant J, Griffin M, Pierce RL, Celi LA. From quality improvement to equality improvement projects: a scoping review and framework. *iScience*. 2023;26(10):107924. https://doi.org/10.1016/j.isci.2023.107924.