HOLOGIC®

Reducing Bias in Breast Cancer Detection AI

Assessing the performance of Hologic's Genius AI™ Detection Solution[7,8] across different race and ethnicities

Ashwini Kshirsagar, Ph.D., Director, Product Owner AI, Research and Development, Hologic, Inc.
Andrew Smith, Ph.D., Vice President, Image Research Breast Health, Hologic, Inc.

Role of AI in screening mammography: Regular screening for breast cancer has been shown to reduce the mortality of breast cancer.[1,2,3] These studies were conducted using 2D imaging technologies. More recently, the utilisation of breast tomosynthesis in screening has been shown to both increase the detection of invasive breast cancer and to reduce false positives compared to older 2D breast imaging.[4,5]

The review of breast tomosynthesis images involves the evaluation of large numbers of thin cross-sectional slices through the breast, with the interpreting physician searching for various indications of cancer including occasionally subtle features such as deposits of microcalcifications or low-contrast masses and asymmetries. This review can be time consuming, and can be challenging, especially in breasts with obscuring parenchymal densities.

As an aid in the review of the images, equipment manufacturers have developed computer-aided detection (CADe) and computer-aided diagnosis (CADx) Artificial Intelligence (AI) systems, which search the image sets and mark suspicious areas and lesions. As a further aid, some of these systems have recently incorporated workflow enhancement tools such as automated lesion correlations, that identify a given marked lesion in one mammographic view (e.g., CC) with its corresponding lesion mark in a different view (e.g., MLO).

AI has been used in breast cancer screening for many years.[6] It operates on both 2D and tomosynthesis images. Referring specifically to CAD systems operating on tomosynthesis datasets, the performance of 3D™ AI (CAD) has been evaluated in a general population in US women.[7] AI is also used in the generation of synthesised 2D and 6-mm images, and in automated breast density determinations. Thus, AI is intertwined in many aspects of breast imaging.

Due to the use of AI in many of its imaging product lines, Hologic makes a concerted effort to collect diverse ethnic and racial data as part of the AI algorithm developments, and to measure the behavior of Hologic products that rely on AI for their performance. The objective of this paper is to describe the strategy to minimise potential bias in AI performance, and to assess the stratified cancer detection performance of Hologic's 3D™ AI (CAD) product, Genius AI Detection solution,[7,8] using images from subgroups of various race and ethnicities.

The need for unbiased AI: AI algorithms can exhibit racial disparities in a wide variety of applications. Some representative examples from the literature are summarised here. Koenecke et al. found that automated speech recognition systems perform poorly for African American speakers, compared to white speakers, and present them with burdens in using these commonly used tools.[9] Adamson and Smith showed how AI in dermatology performed more poorly in people with darker skin types, and that this limitation could marginalise some groups, and delay the widespread adoption of a very promising technology.[10] Obermeyer et al. discussed the issue of a widely-used medical risk algorithm that scored black patients identically to white patients, despite their being considerably sicker.[11] This disparity might reduce the needed medical care provided to this patient group. Seyyed-Kalantari et al. described how AI-trained models on large public chest X-ray datasets displayed systematic bias among differing racial, age, and sex subgroups, which could mean that individuals would be given incorrect diagnoses at a greater extent than other subgroups.[12]

It is Hologic's goal to minimise such disparities in its breast imaging AI applications. AI breast cancer screening products are used in the diagnostic evaluation of patients in screening populations, such as the detection of breast cancer. AI also has ancillary roles such as in risk prediction and allocation of resources, an example of which is the determination of breast density and the suitability of secondary screening such as ultrasound. Therefore, it is important to ensure that the performance of AI results in equity in health care for patients of varying ethnicities and racial groups.[13]

Is bias in an AI algorithm, as defined by performance variations in differing racial and ethnic subgroups, likely? There are racial and ethnic differences reflected in certain breast imaging biomarkers, such as breast density[14] and body habitus,[15] and, perhaps surprisingly, mammograms exhibit other not as clearly understood differences among differing subgroups. For example, the ability of AI to detect a patient's race in a mammogram with an accuracy better than chance has been demonstrated.[16,17] This performance persists even accounting for the ethnicity biomarker variations such as having breast densities that vary from the average. Therefore, given that some imaging biomarkers have the potential to vary among racial and ethnic subgroups, it is important to establish a strategy to minimise potential bias and to ascertain the performance of AI in these population subgroups.

Strategies to reduce bias: The creation of artificial intelligence algorithms starts with the collection of large, high-quality databases of patient information and images. These databases are used in developing and training the algorithms, and in testing the algorithms to ensure they meet expected performance. It is important that, as much as possible, the database reflects the diversity of racial and ethnic groups in the population, so that the developed algorithms do not unexpectedly perform sub-optimally in some of these groups.

The possibility that AI can identify a patient's race in a mammogram implies that the AI algorithm may train and perform differently in different racial groups, and therefore the database used in AI development should contain examples of varying subgroups so that the algorithm gets the opportunity to maximise performance across all subgroups.

AI development relies heavily on the quality and quantity of training data. Large quantities of data are especially important in deep learning approaches compared with older machine learning technologies. AI systems are designed to learn and make decisions based on patterns and information present in the data they are exposed to during training. During the training process, an AI model is fed large amounts of data, which then iteratively refines its internal parameters to recognise patterns and correlations. The model adjusts its behavior based on the information present in the training data, making it crucial to ensure that the data is diverse, representative, and reflective of the real-world scenarios the AI system will encounter. The quality of training data directly impacts the performance and reliability of AI models.[18] If the data is biased, incomplete, or lacks diversity, the AI system may exhibit unintended behaviors. Therefore, the role of training data goes beyond merely providing information – it serves as the foundation for the ethical and unbiased functioning of AI systems. To achieve this, the training data must mirror the diversity and complexity of the environments in which the AI system will operate. A narrow or biased dataset may result in an AI model that fails to generalise well, leading to poor performance in real-world applications. The quality, diversity, and representativeness of the data directly influence the capabilities and ethical considerations of AI systems. Recognising the importance of diverse training data is a crucial step toward building AI models that are robust, fair, and capable of making accurate decisions in various real-world scenarios.

In the field of screening mammography AI, the role of training data takes on added significance due to the importance of breast cancer detection. Hologic products that are powered by AI algorithms are built using a robust dataset supported by solid ground truth. Ground truth is established using confirmed pathological outcomes for abnormal patient cases and using temporal follow-up for normal patient cases. Truth marking of the sites of biopsies is established using post-biopsy clip-placement images. The following sections are intended to demonstrate the foundation of this robust dataset that encompasses a wide range of demographics, crucial to ensuring that AI models can provide equitable and effective breast cancer detection. While this paper will focus on the performance of the Genius AI™ Detection product line, the image database and associated clinical database illustrated in this work also serves as a foundation of other AI-containing Hologic products including C-View™, Intelligent 2D™, 3DQuorum™ and Quantra™ technologies. Hologic makes significant investments to create racially and ethnically diverse datasets to avoid disparities in diagnosis, and Hologic is committed to the use of comprehensive and representative training data to enhance the reliability and equitability of all AI-powered products in Hologic's portfolio.

Hologic's diverse AI database: Hologic has assembled a racially and ethnically diverse database for use in AI development, with the explicit goal of preventing algorithmic racial and ethnicity bias in its products. This required collecting images and patient information from multiple breast imaging centers, ensuring that the database contained a heterogeneous population representative of the major racial groups in the US, as described in the US Population Census.[19] The data are collected under Institutional Review Board approvals with waiver of consent and the data are fully anonymised per HIPAA standards to avoid any selection bias.[20] In addition to the fundamental imaging data required for training AI models, race and ethnicity data was collected, when available, from the mammography facilities that Hologic partners with on data collection. Hologic's goal is to follow the AI training guiding principles following known methodologies, e.g. see Alexander et al, to understand and address the potential impact of the AI training database on the performance of its products in different demographic groups.[21]

Race and ethnicity constitution of Hologic's AI training database: The diversity of Hologic's database was assessed by analysing all available race and ethnicity information data that was gathered from the imaging facilities. Race and ethnicity information was not available for every patient; however, the data that is available shows diversity in the training database as illustrated below. The analysis in this paper was performed on an approximately 50,000 patient subset of the training database for which racial and ethnic information was available. Approximately 4% of patients amongst this subset declined to specify either race or ethnicity information and were eliminated from the analysis.

The following racial and ethnic categories were used in our database segmentation:

Race

- White
- African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Other Pacific Islander
- Other Race
- Multi-race

Ethnicity

- Hispanic or Latino
- Non-Hispanic or Latino

Each patient separately self-identified themselves into these categories. Hologic analysed the racial and ethnic makeup of its database by stratifying on categories employed in the US Population Census, as described by Jensen et al.[22] They suggested combining race and ethnicity information to create groups by using mutually exclusive racial and ethnic (nonoverlapping) categories. They indicate that that people of Hispanic origin may be of any race and proposed to stratify using the following categories:

- Hispanic
- White alone (non-Hispanic)
- African American alone (non-Hispanic)
- American Indian and Alaska Native alone (non-Hispanic)
- Asian alone (non-Hispanic)
- Native Hawaiian and Other Pacific Islander alone (non-Hispanic)
- Some Other Race alone (non-Hispanic)
- Multiracial (non-Hispanic)

The distribution of the ethnic and racial breakdown of the patients in the evaluated database in these categories can be seen in Figure 1.

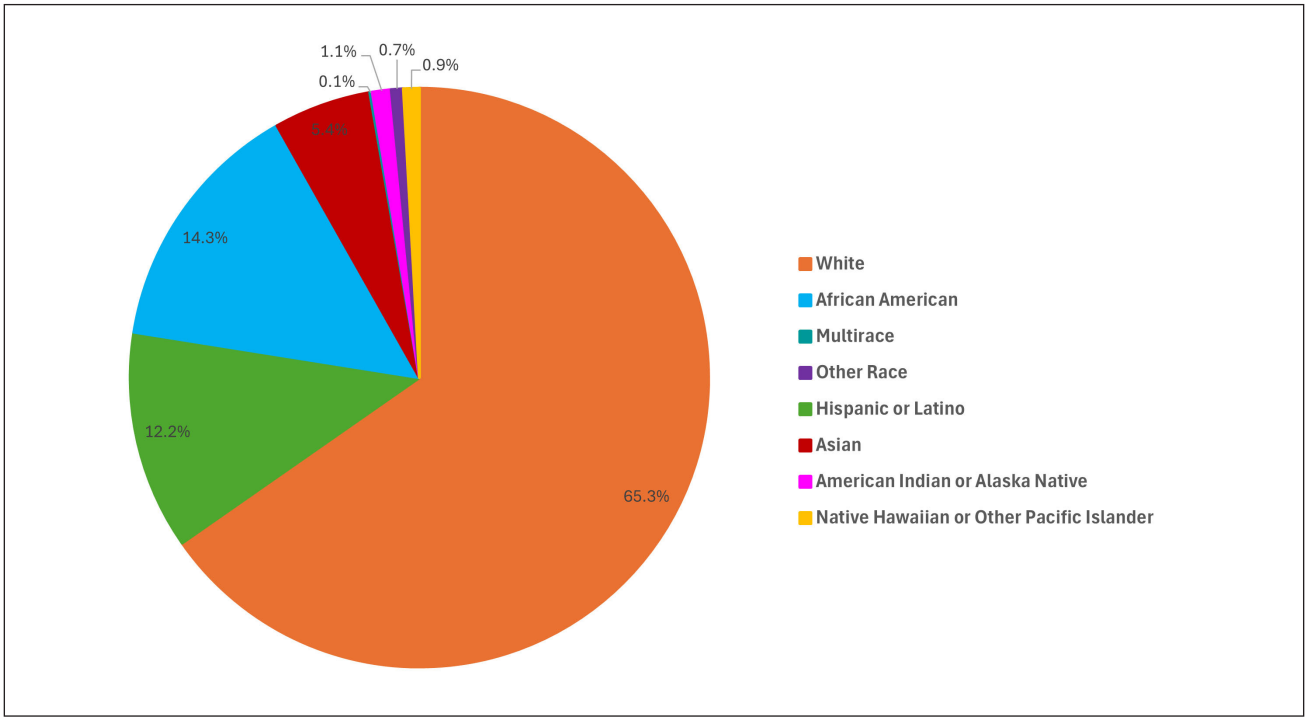

Figure 1. Distribution of racial and ethnic categories in Hologic AI database.

For comparison, Figure 2 shows data from the 2020 US Census, using the same categories.[23] A comparison of these two figures shows that Hologic has assembled a database broadly representative of racial groups in the US.

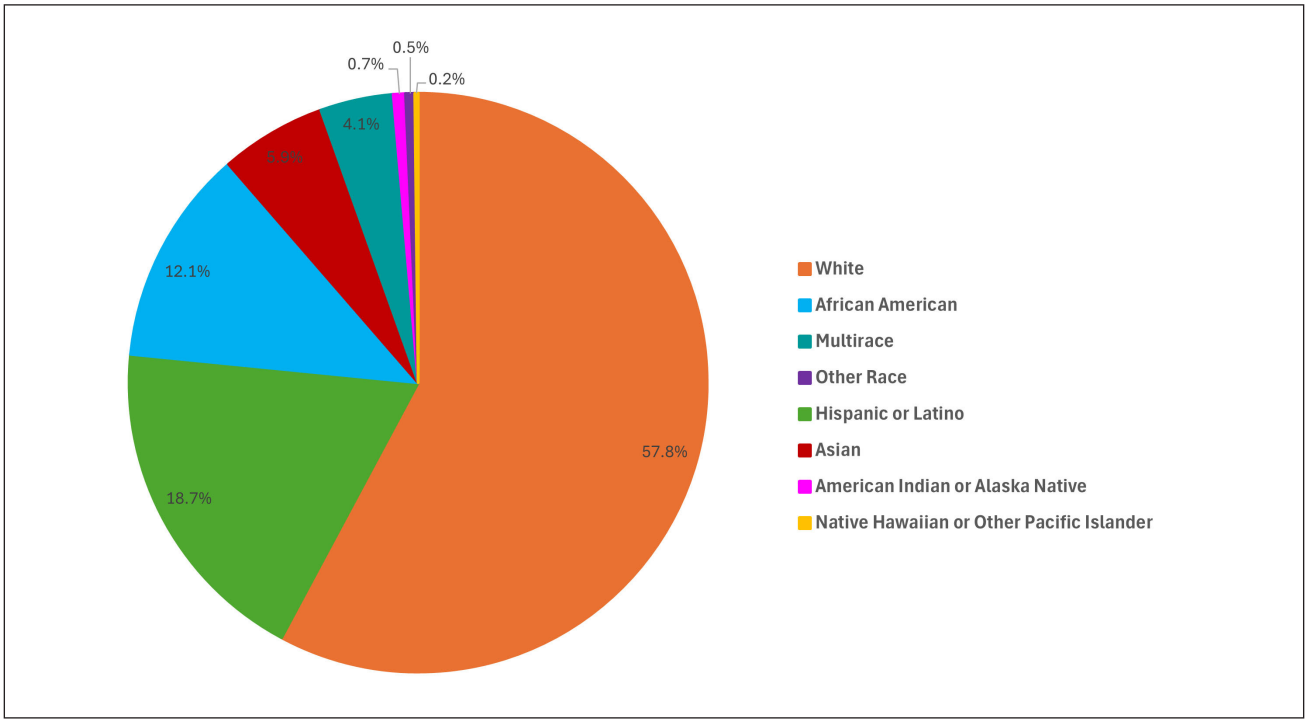

Figure 2. US Population from 2020 Census, stratified by race and ethnicity

Performance of Genius AI™ Detection solution[7,8] in a diverse population: The overall high performance of Hologic's Genius AI Detection solution has been previously reported, with radiologists demonstrating an improvement in clinical accuracy (improved Area under the ROC curve) when interpreting tomosynthesis images using Genius AI Detection solution as compared to not using it, and a +9% in cancer sensitivity.[24]

The cancer detection performance, as a function of racial and ethnic subgroups, of Hologic's Genius AI Detection solution (version 2.0) was analysed using a subset of the AI database. For this evaluation, approximately 8,000 cases containing known racial and ethnic and ground truth data were consecutively selected from the database from a pool of biopsy-proven malignant cancer cases and negative cases that were read as BI-RADS 1 or 2 at screening. It is important to note that these cases were not involved in the training and development processes for the Genius AI Detection solution used in this evaluation. These cases were selected based on the availability of complete imaging data and information about race and ethnicity from the curated data pool available; no other selection criteria were implemented. The objective was to measure and compare the Genius AI Detection solution performance in the various racial and ethnic subgroups. Because of the relatively smaller populations in the racial subgroups of American Indians, Alaskan Natives, and Native Hawaiian or Other Pacific Islanders, for the purpose of this data shown here these subgroups have been grouped into a single category Native & Pacific Islanders.

The performance of the Genius AI Detection solution was measured using several metrics: Receiver Operating Characteristics (ROC) curves, Area under the ROC curve (AUROC) and cancer detection (sensitivity). ROC curves and their areas and uncertainties were calculated using a maximum likelihood estimation of binormal ROC curves from continuously distributed test results.[25]

The ROC curves in Figure 3 show the performance of the Genius AI Detection solution as a function of sensitivity and specificity. The ROC curves demonstrate similar cancer detection performance for the different racial groups.

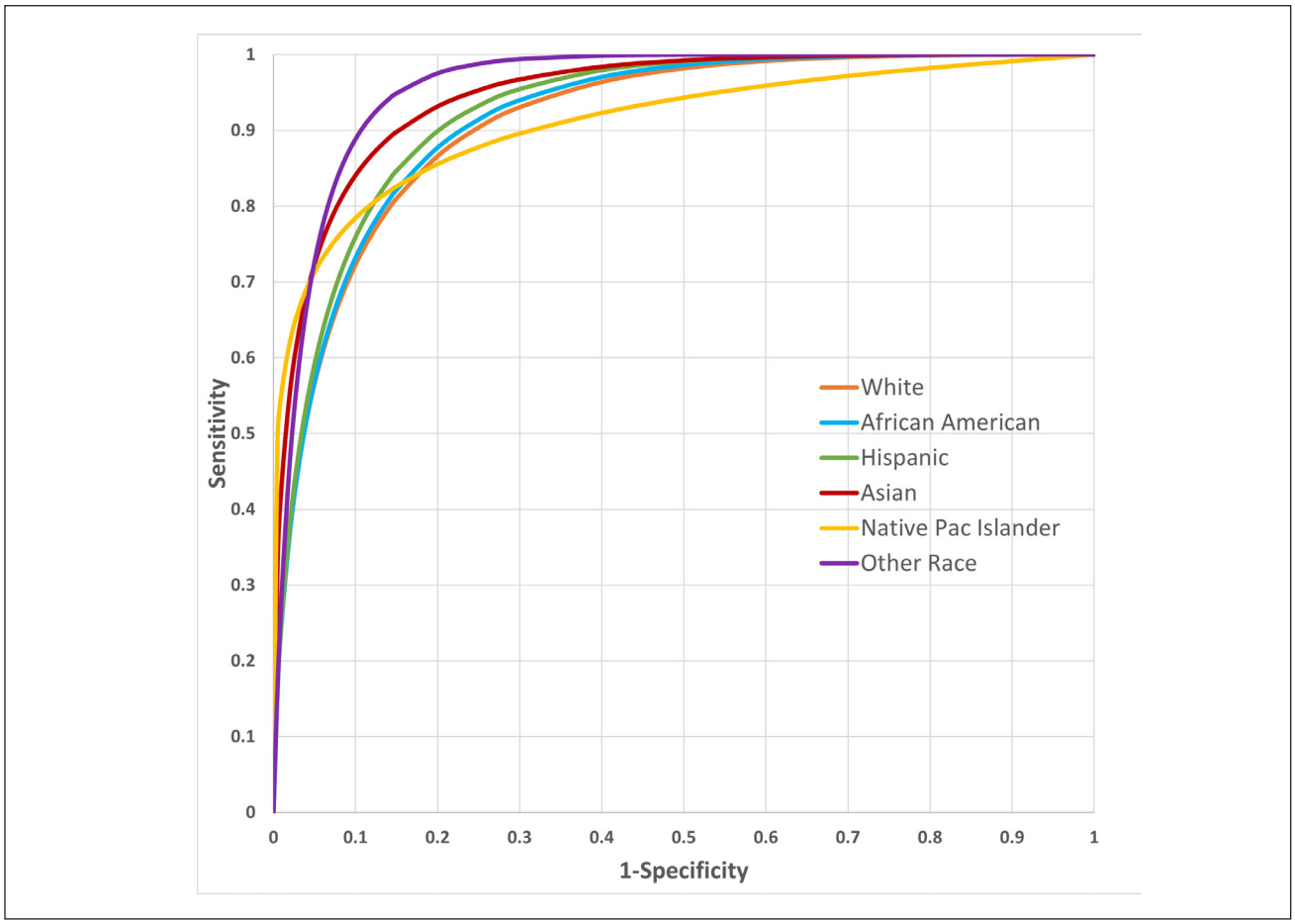

Figure 3. ROC curves, stratified by race and ethnicity

The area under the ROC curve, AUROC, is an overall measure of the performance of Genius AI™ Detection solution. The AUROC results, including the uncertainty, are shown in Figure 4. To within the uncertainty, the AUROC for each racial subgroup have nearly identical performance. To the extent possible given the error bars in the results, Genius AI Detection solution has been shown to have similar performance across a variety of racial groups found in the diverse US population and across the globe.

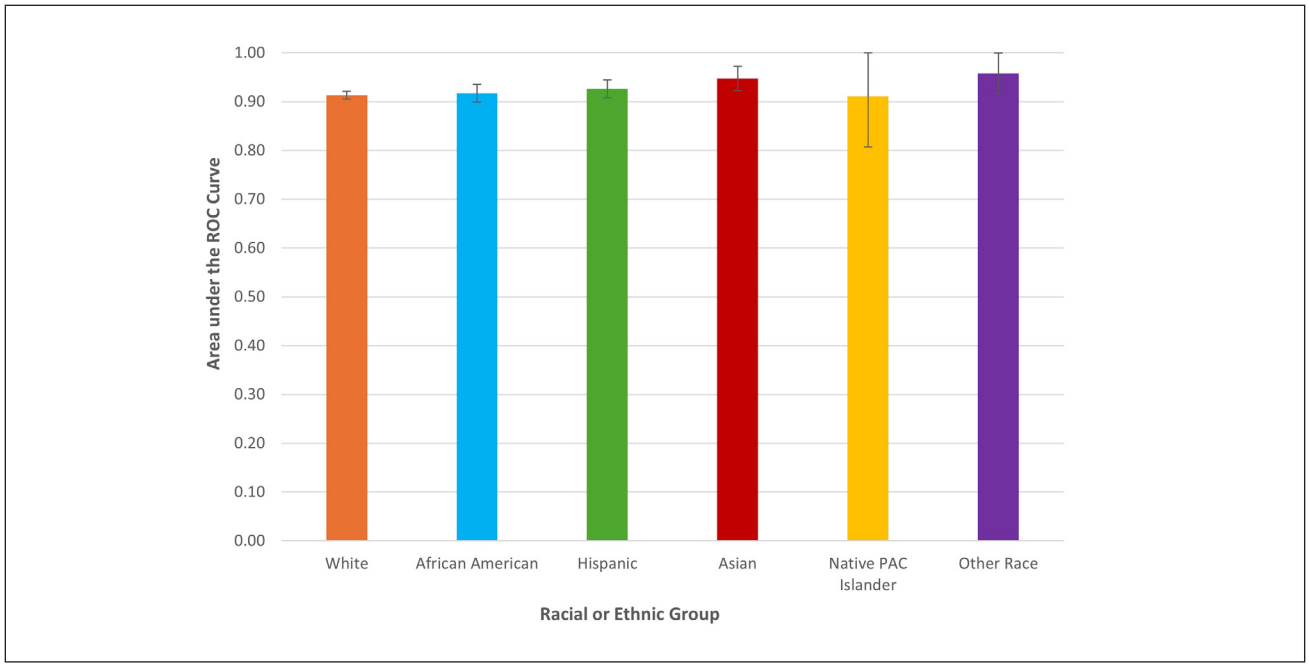

Figure 4. Area under the ROC Curve for the racial and ethnic subgroups (Error bars indicate 95% confidence intervals)

Cancer detection performance of the Genius AI Detection solution was measured by determining the percentage of known cancers in the testing database that were properly identified. Cancer in a given patient was considered as detected if a Genius AI Detection solution mark was accurately placed on the location of cancer on at least one screening image view. The cancer detection performance of Genius AI Detection solution in terms of racial subgroups is shown in Figure 5. Genius AI Detection solution performed similarly in each racial subgroup.

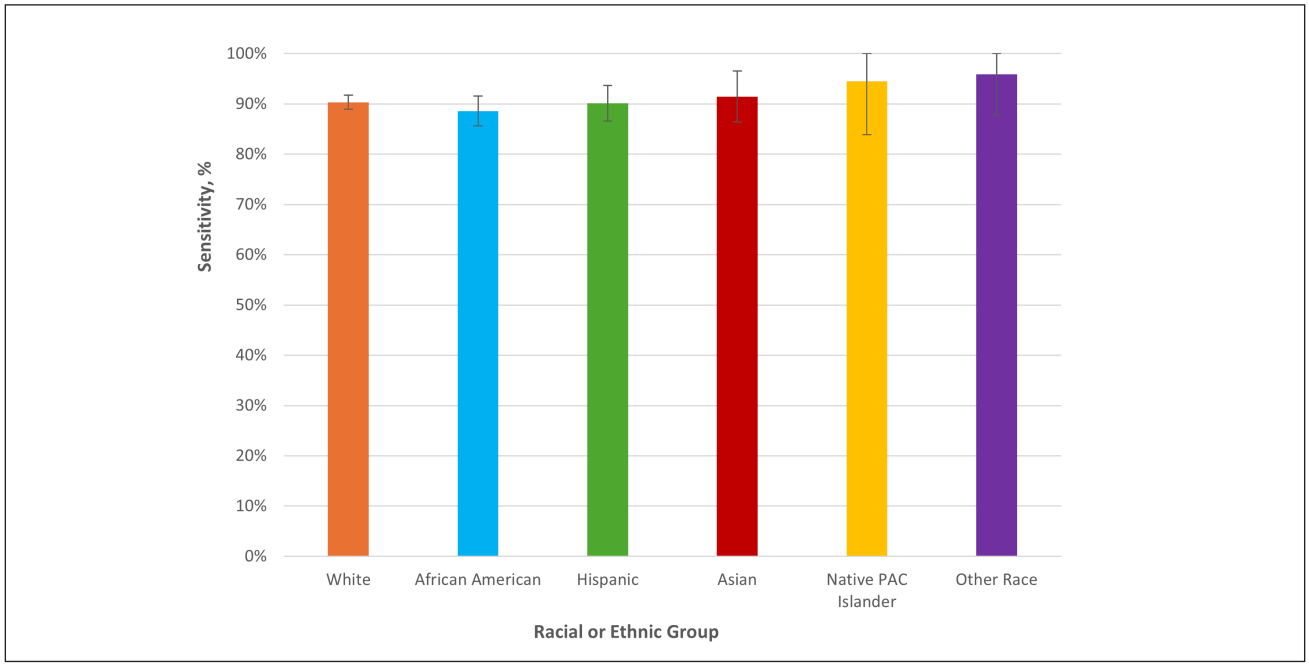

Figure 5. Cancer detection sensitivity as a function of racial subgroup (Error bars indicate 95% confidence intervals)

Conclusion: The database of patient information collected by Hologic for use in its AI product lines, represents a heterogeneous mix of subjects with a variety of races and ethnicities that is broadly representative of their distribution in the US. While not explicitly tested in patient data collected from outside the US, because these races and ethnicities are also found across the world, the AI-based product might be expected to behave similarly in these tested racial and ethnic populations in countries outside of the US.

The effectiveness of the diversity of the database was tested using the product Genius AI™ Detection solution (version 2.0). The performance of Genius AI Detection solution was measured in cohorts representing differing racial groups. The observed performance, as measured using the area under the ROC curve (AUROC) which provides a comprehensive measure of the balance between sensitivity and specificity, was similar for each of the racial groups and also similar to the overall average performance of the entire population. In addition, the cancer detection sensitivity was also similar across all the racial groups analysed. While a larger systematic study focused on the minority population would be desirable, to improve statistical power in the rarer subgroups, the data presented here indicates that Genius AI Detection solution offers similar performance across the racial and ethnic subgroups tested.

The effort that Hologic has made in collecting its diverse AI database reflects the larger message of its commitment to improving the health of women worldwide.

Acknowledgments: We would like to acknowledge the help from Carol Fisk, Christine Jerome, and Kathleen Willison for assistance with data collection and Dhruv Chamania for assistance with data management infrastructure.

Breast & Skeletal Health | Hologic.co.uk | euinfo@hologic.com

CE 2797 EC REP Hologic BV, Da Vincilaan 5, 1930 Zaventem, Belgium.

References: 1. Smith, Robert A., et al. "The randomised trials of breast cancer screening: what have we learned?." Radiologic Clinics 42.5 (2004): 793-806. **2.** Hendrick, R. Edward, et al. "Benefit of screening mammography in women aged 40-49: a new meta-analysis of randomised controlled trials." JNCI Monographs 1997.22 (1997): 87-92. **3.** Tabar L, Vitak B, Tony HH, Yen MF, Duffy SW, Smith RA. Beyond randomised controlled trials: organised mammographic screening substantially reduces breast carcinoma mortality. Cancer 2001;91:1724-31. **4.** Friedewald, Sarah M., et al. "Breast cancer screening using tomosynthesis in combination with digital mammography." Jama 311.24 (2014): 2499-2507. **5.** Conant, Emily F., et al. "Five consecutive years of screening with digital breast tomosynthesis: outcomes by screening year and round." Radiology 295.2 (2020): 285-293. **6.** Image Checker CAD for Mammography. US FDA PMA P970058 (1998-2012). **7.** Genius AI Detection 2.0 with CC-MLO Correlation. US FDA 510(k) K230096 (2023). **8.** Genius AI Detection solution. US FDA 510(k) K221449 (2020). **9.** Koenecke, Allison, et al. "Racial disparities in automated speech recognition." Proceedings of the National Academy of Sciences 117.14 (2020): 7684-7689. **10.** Adamson, Adewole S., and Avery Smith. "Machine learning and health care disparities in dermatology." JAMA dermatology 154.11 (2018): 1247-1248. **11.** Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366.6464 (2019): 447-453. **12.** Seyyed-Kalantari, Laleh, et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers." BIOCOMPUTING 2021: proceedings of the Pacific symposium. 2020. **13.** Chin, Marshall H., et al. "Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care." JAMA Network Open 6.12 (2023): e2345050-e2345050. **14.** del Carmen, Marcela G., et al. "Mammographic breast density and race." American Journal of Roentgenology 188.4 (2007): 1147-1150. **15.** Bissell, Michael CS, et al. "Breast cancer population attributable risk proportions associated with body mass index and breast density by race/ethnicity and menopausal status." Cancer Epidemiology, Biomarkers & Prevention 29.10 (2020): 2048-2056. **16.** Gichoya, Judy Wawira, et al. "AI recognition of patient race in medical imaging: a modelling study." The Lancet Digital Health 4.6 (2022): e406-e414. **17.** Banerjee, Imon, et al. "Reading race: AI recognises patient's racial identity in medical images." arXiv preprint arXiv:2107.10356 (2021). **18.** Gupta, N., Mujumdar, S., Patel, H., et al, Data quality for machine learning tasks. Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 4040-4041), 2021. **19.** Jensen, Eric et al. "Measuring Racial and Ethnic Diversity for the 2020 Census." United States Census Bureau (2021). https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html. **20.** Gichoya JW, Thomas K, Celi LA, et al. AI pitfalls and what not to do: mitigating bias in AI. Br J Radiol (2023) 10.1259/bjr.20230023. **21.** Chin, Marshall H., et al. "Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care." JAMA Network Open 6.12 (2023): e2345050-e2345050. **22.** Jensen, Eric et al. "Measuring Racial and Ethnic Diversity for the 2020 Census." United States Census Bureau (2021). https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html. **23.** United States Census Bureau, cited 3/26/2024. https://data.census.gov/table/DECENNIALDP2020.DP1?g=010XX00US&d=DEC%20Demographic%20Profile. **24.** Genius AI Detection for Breast Tomosynthesis, cited 3/27/2024. https://www.hologic.com/sites/default/files/2020_12/WP-00178_Rev02_GeniusAI_Detection-white-paper-6979r10p.pdf. **25.** Eng J. ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University [updated 2014 March 19; cited 2/2/2024]. Available from: http://www.jrocfit.org.