Emory Breast Imaging Dataset (EMBED) v2 – a Racially Diverse, Multi-modal Dataset of 1.2M Breast Imaging Exams and Associated Histopathology

Rohan Satya Isaac, MS, Senior Data Analayst, School of Medicine, Emory University
Beatrice Brown-Mulry; Aawez Mansuri, MS; Theo Dapamede, MD, PhD; Chad Robichaux, MPH; Frank Li, PhD; Mohammedreza Chavoshi, MD; Judy Gichoya, MS, MD, FSIIM; Hari Trivedi, MD

Introduction

We developed the EMory BrEast Imaging Datset (EMBED) in 2023, which is currently used at over 300 institutions worldwide and has been used to train or validate multiple FDA-cleared models. This dataset contained 350,000 2D and synthetic-2D mammograms along with patient demographics, risk factors, and pathologic outcomes for 116,000 patients. Since then, we have expanded EMBED with four additional years of data and added digital breast tomosynthesis (DBT), US, and MRI modalities, to create EMBEDv2.

Hypothesis

Substantial expansion of EMBED will enhance its utility for training and validating advanced breast imaging models, improving generalizability and performance across diverse patient populations.

Methods

We queried Magview software (Fulton, MD) for net new breast imaging exams since December, 2020. Relevant patient demographics, pathology reports, receptor, and recurrence information were extracted from the EHR. Discrepant data, such as changes in patient ID, breast density, and pathologic outcomes were manually reviewed and resolved. Outcomes are harmonized with the Georgia Department of Public Health Cancer Registry. Enrichment steps for both clinical and metadata were performed similar to EMBEDv1. Digital histopathology was aggregated for previously digitized patients, and will be collected prospectively beginning October, 2024.

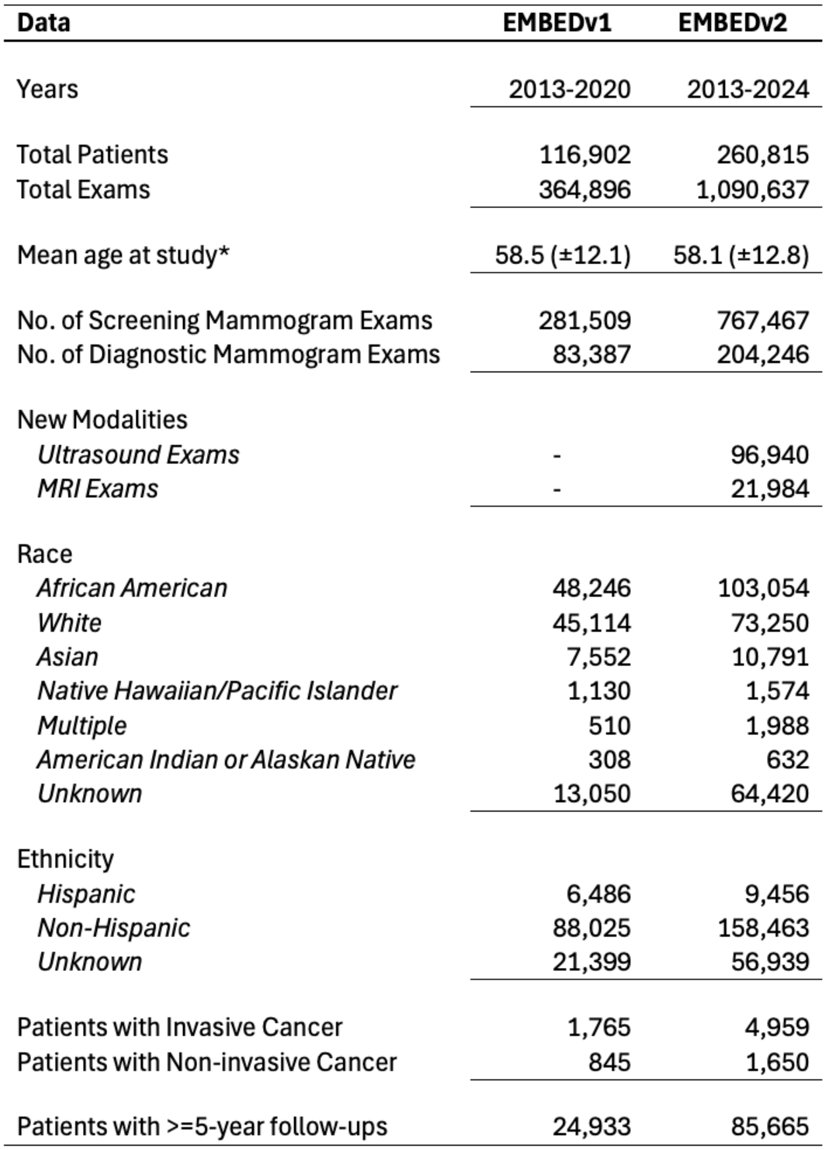
Results

EMBEDv2 now encompasses 2013-2024 and has expanded from 116,177 to 260,815 patients and from 383,421 to 1,090,637 exams (Table 1). The dataset contains 103,054 (40.3%) African American, 73,250 (28.6%) White, and 9,456 (4.2%) Hispanic patients. There are 767,500 (70.4%) screening mammograms, 204,246 (18.7%) diagnostic mammograms, 96,940 (8.9%) US, and 21,984 (2.0%) MRI exams. Ground truth pathologic outcomes are available for all biopsied patients with 4,959 (1.9%) invasive and 1,650 (0.6%) non-invasive cancers. 5 years of follow-up is available for 85,665 (32.8%) patients.

Conclusion

EMBED V2 contains 1.2M exams from 2020-2024, including two additional clinical sites, and expands to include digital breast tomosynthesis (DBT), US, and MRI exams. A subset of the dataset will again be released for researchers and be made available for commercial use.

Figure(s)



Data	EMBEDv1	EMBEDv2
Years	2013-2020	2013-2024
Total Patients	116,902	260,815
Total Exams	364,896	1,090,637
Mean age at study*	58.5 (±12.1)	58.1 (±12.8)
No. of Screening Mammogram Exams	281,509	767,467
No. of Diagnostic Mammogram Exams	83,387	204,246
New Modalities		
Ultrasound Exams	-	96,940
MRI Exams	-	21,984
Race		
African American	48,246	103,054
White	45,114	73,250
Asian	7,552	10,791
Native Hawaiian/Pacific Islander	1,130	1,574
Multiple	510	1,988
American Indian or Alaskan Native	308	632
Unknown	13,050	64,420
Ethnicity		
Hispanic	6,486	9,456
Non-Hispanic	88,025	158,463
Unknown	21,399	56,939
Patients with Invasive Cancer	1,765	4,959
Patients with Non-invasive Cancer	845	1,650
Patients with >=5-year follow-ups	24,933	85,665

Table 1. Comparison of characteristics between EMBED v1 and EMBED v2. * Data means ± SDs.

Keywords

Artificial Intelligence/Machine Learning; Enterprise Imaging; Imaging Research