

Do Deep Embedding-Derived Imaging Phenotypes Reveal Breast Cancer Risk Heterogeneity Beyond Scalar Risk Stratification?

Lydia J. Sollis^{1,2} Yurii Shvetsov¹ Brandon Quon¹ Thomas K. Wolfgruber¹ Arianna Bunnell^{1,2} Brenda Y. Hernandez¹ John A. Shepherd¹

¹University of Hawai'i Cancer Center • ²University of Hawai'i at Mānoa

IWBI 2026 • International Workshop on Breast Imaging • Thessaloniki, Greece

UNIVERSITY OF HAWAII
CANCER CENTER

1 Background & Purpose

Deep learning (DL) mammography risk models such as *Mirai* [1] output only scalar breast cancer (BC) risk estimates, leaving their high-dimensional representations uninterpreted and obscuring heterogeneity among women with similar predicted risk.

Unsupervised clustering of mammographic features has been proposed to identify imaging phenotypes beyond conventional risk factors. Kontos et al. [2] clustered radiomic texture features in 2,029 women, finding four parenchymal-complexity phenotypes associated with BC independent of density (AUC 0.84 vs 0.80); but they did not test whether these added structure beyond their own complexity score, and the complexity–cancer relationship was non-monotonic. Winham et al. [3] extended this to ~30,000 women and 390 features, identifying six clusters and six principal components independently associated with invasive BC, capturing structure along multiple risk-relevant axes. Separately, Xu et al. [4] clustered Mirai's patch-level encoder embeddings, recovering distinct tissue phenotypes (dense fibroglandular tissue, architectural distortion, microcalcifications) that exam-level representations may discard during attention-based compression.

Whether Mirai's post-attention exam-level embeddings retain such phenotypic richness remains untested. More broadly, stable clusters from any continuous representation will inevitably differ across correlated variables, but this alone does not prove the presence of genuinely distinct phenotypes rather than arbitrary partitions along the representation's primary axis of differentiation.

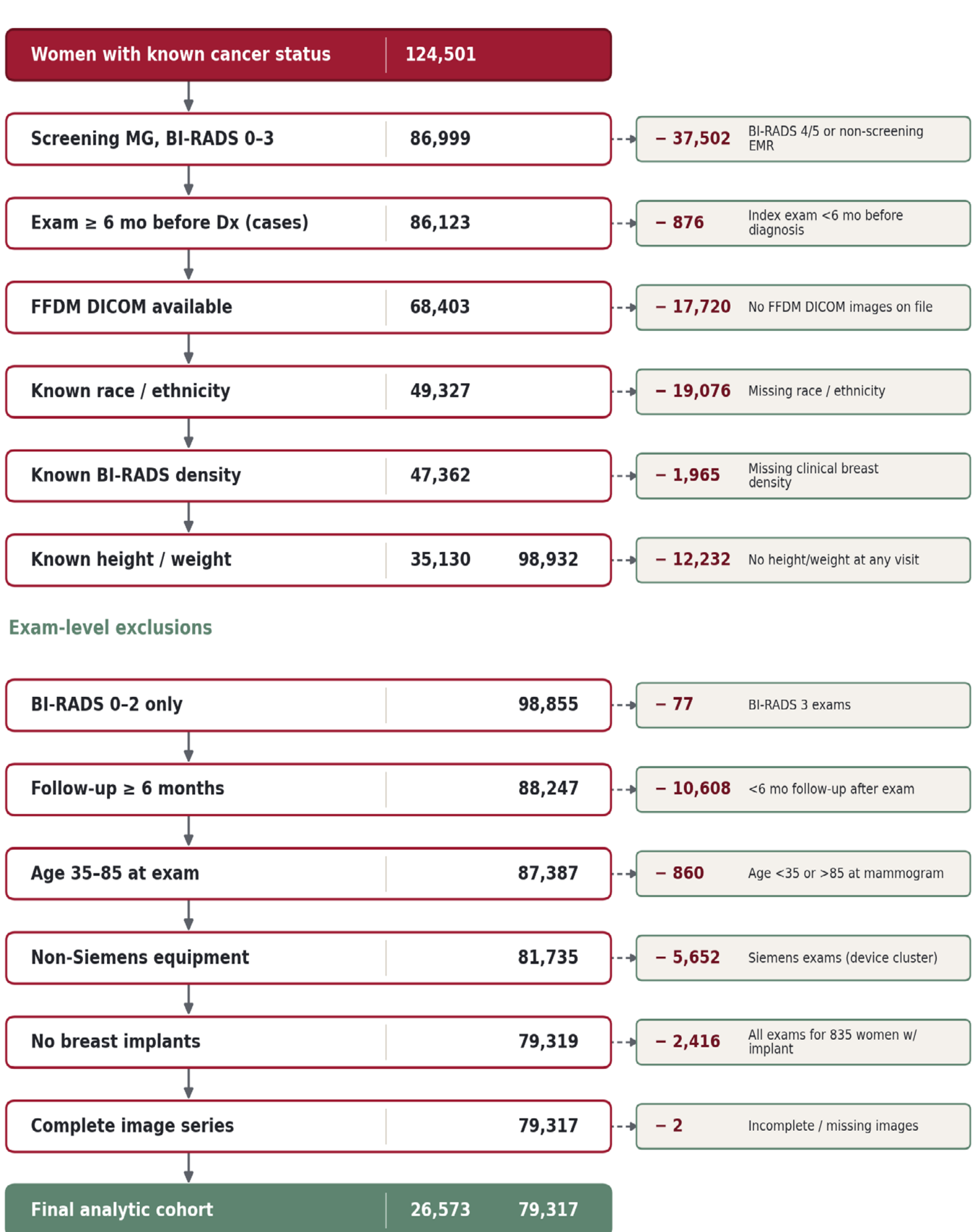
We test whether unsupervised clustering of Mirai's exam-level embeddings recovers **stable, clinically meaningful phenotypes beyond scalar risk**, benchmarking against risk-ordered control partitions as a negative control.

2 Methods

Study Cohort Derivation

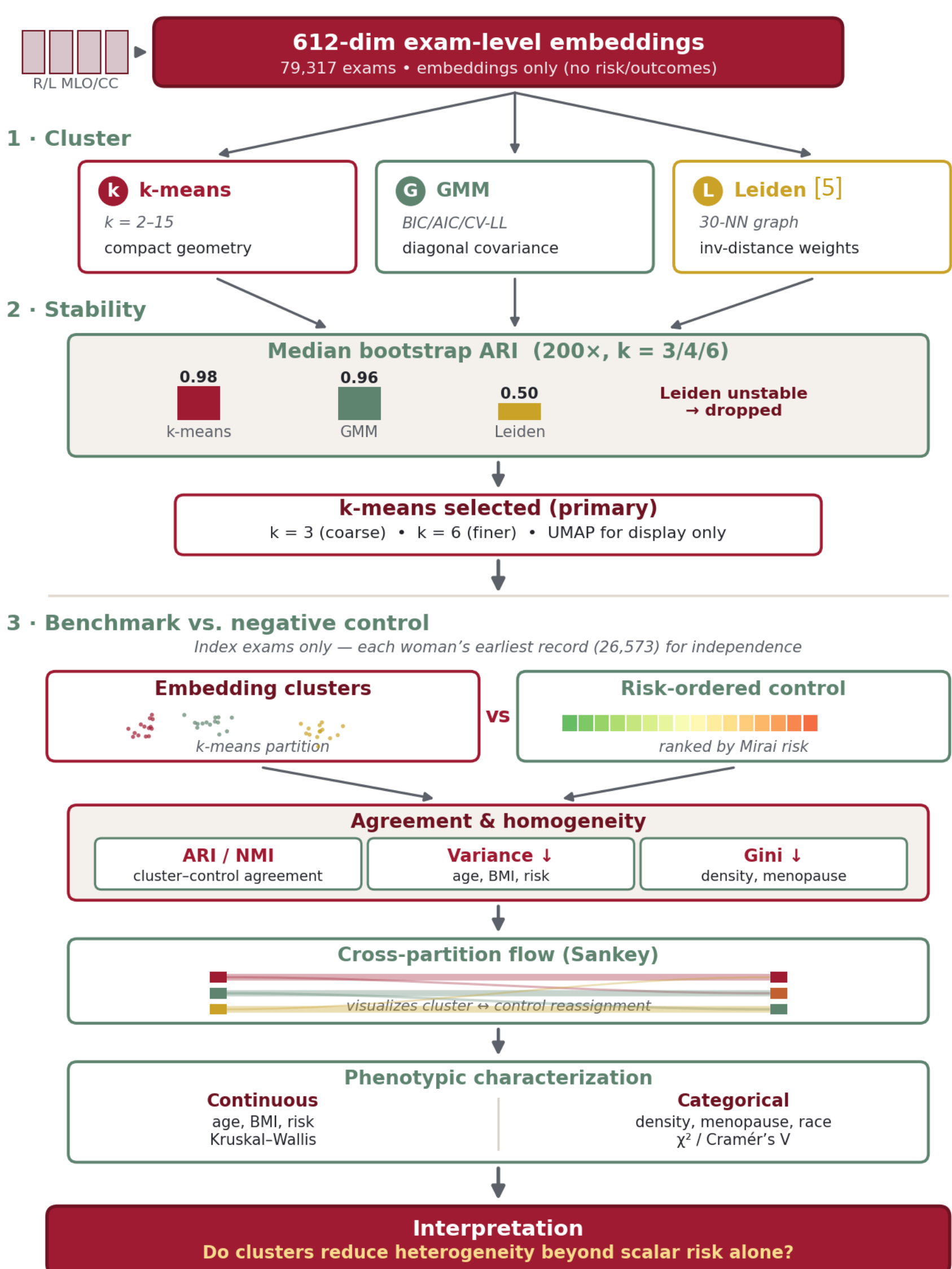
Hawai'i and Pacific Islands Mammography Registry (HIPIMR), screening mammograms 2010–2020, linked to the Hawai'i Tumor Registry (HTR). Funding: NIH R01CA263491, U54CA143728.

Woman-level eligibility



Analysis Pipeline

Clustering → stability → statistical benchmarking



3 Results

Table 1. Cohort characteristics for the 26,573 women (79,317 screening examinations) in the HIPIMR analytic cohort, 2010–2020, stratified by breast cancer status during follow-up.

Table 1. Cohort Characteristics			
Characteristic	Overall	No BC in follow-up	BC in follow-up
N women	26,573	25,482	1,091
N examinations	79,317	76,402	2,915
Median exams per woman (IQR)	2 (1–4)	2 (1–4)	3 (2–4)
Age at index exam, mean ± SD (y)	55.0 ± 11.0	54.8 ± 11.0	58.3 ± 10.7
Median Mirai 5-yr risk (IQR)	1.8% (1.5–2.0)	1.8% (1.5–2.0)	1.9% (1.6–2.1)
Race / ethnicity, n (%)			
Japanese American	7,385 (27.8)	6,991 (27.4)	394 (36.1)
Filipino/a American	4,793 (18.0)	4,647 (18.2)	146 (13.4)
Native Hawaiian / Pacific Islander	4,642 (17.5)	4,449 (17.5)	193 (17.7)
White	4,062 (15.3)	3,946 (15.5)	116 (10.6)
Chinese American	2,712 (10.2)	2,603 (10.2)	109 (10.0)
Korean American	1,105 (4.2)	1,054 (4.1)	51 (4.7)
Other Asian	929 (3.5)	909 (3.6)	20 (1.8)
Hispanic	465 (1.7)	414 (1.6)	51 (4.7)
BMI, mean ± SD (kg/m ²)	27.1 ± 6.7	27.1 ± 6.7	27.2 ± 6.2
BI-RADS density, n (%)			
A – Almost entirely fatty	1,324 (5.0)	1,287 (5.1)	37 (3.4)
B – Scattered fibroglandular	11,432 (43.0)	10,927 (42.9)	505 (46.3)
C – Heterogeneously dense	11,300 (42.5)	10,846 (42.6)	454 (41.6)
D – Extremely dense	2,517 (9.5)	2,422 (9.5)	95 (8.7)
Dense breasts (C or D)	13,817 (52.0)	13,268 (52.1)	549 (50.3)
Postmenopausal, n (%)	18,073 (68.0)	17,289 (67.8)	784 (71.9)
Median follow-up, y (IQR)	6.5 (3.4–9.0)	6.4 (3.3–8.9)	8.4 (6.0–9.7)
Cancer type, n (%)			
Invasive	–	–	828 (75.9)
DCIS only	–	–	263 (24.1)
Time to diagnosis, median y (IQR)	–	–	4.4 (2.6–6.5)

Data are n (%) unless noted. Race/ethnicity groups 51% shown. BC = breast cancer; DCIS = ductal carcinoma in situ; IQR = interquartile range. Menopause imputed postmenopausal for age > 55 with missing data.

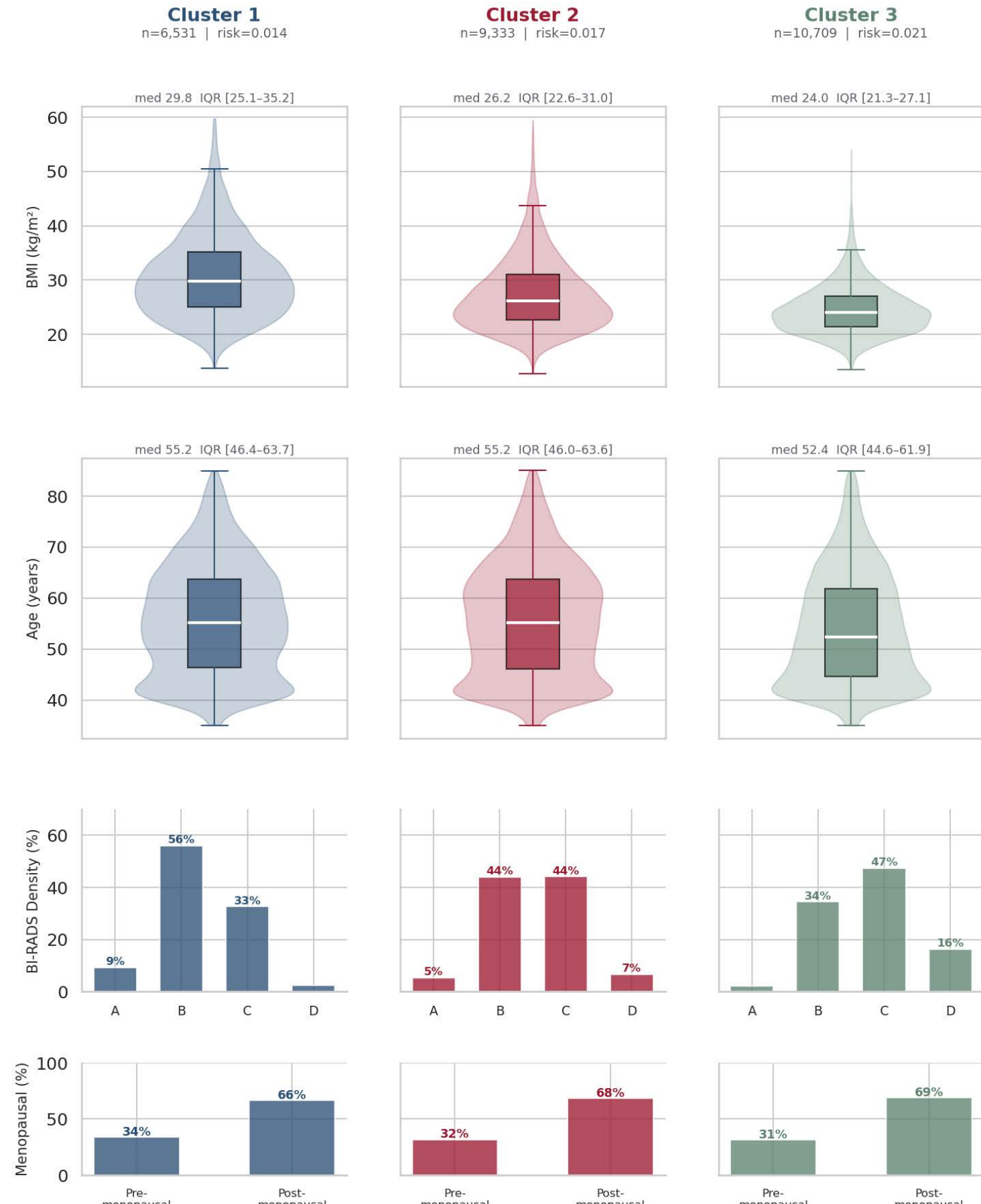


Figure 2. Phenotypic profiles of the embedding-derived clusters (k=3), with columns ordered by mean predicted Mirai 5-year risk (Cluster 1 lowest, Cluster 3 highest). Rows show within-cluster distributions of BMI and age (violin plus box plots), BI-RADS breast density categories A–D, and menopausal status. Higher-risk clusters are characterized by progressively lower BMI and higher mammographic density.

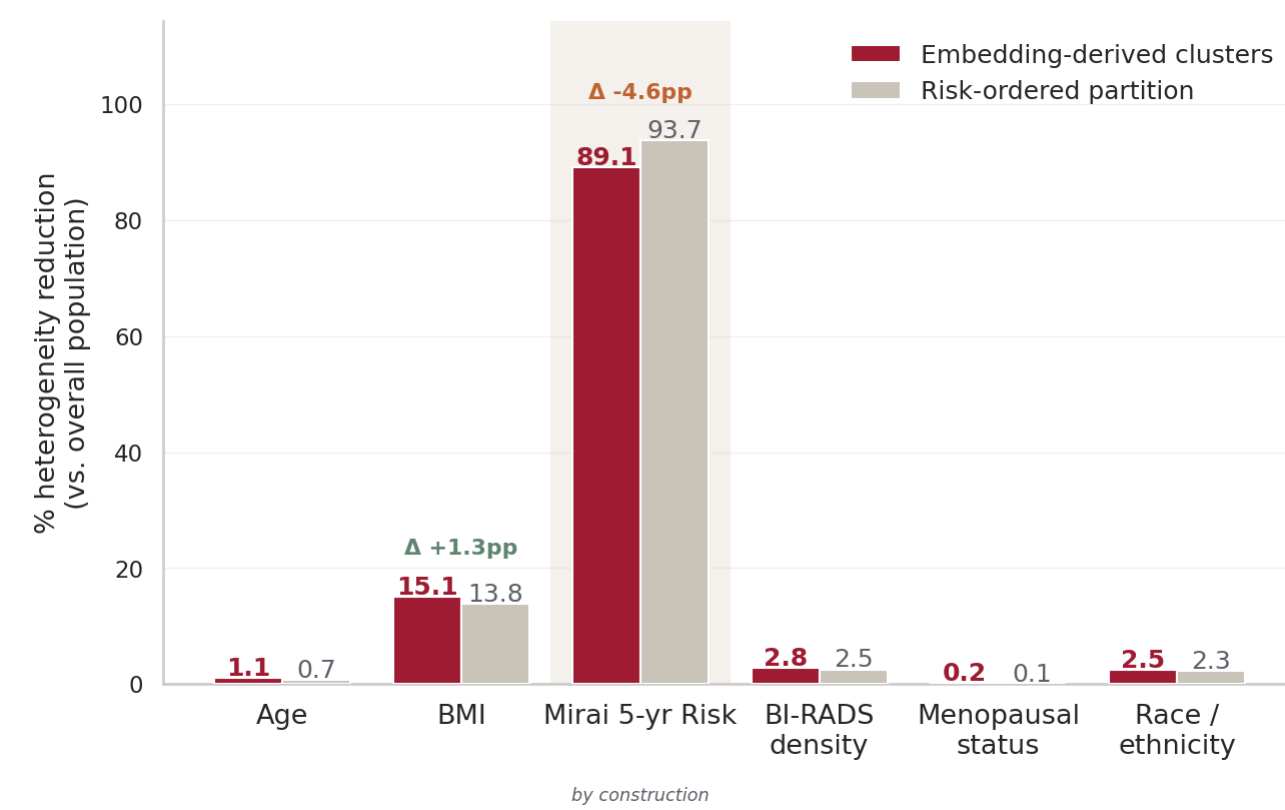


Figure 4. Within-cluster heterogeneity reduction relative to the overall population for embedding-derived clusters versus risk-ordered partitions, across clinical variables (k=3, left; k=6, right). Heterogeneity was quantified by within-cluster variance reduction (continuous variables) and Gini impurity reduction (categorical variables). Embedding-derived clusters showed only a modest advantage in homogeneity over the risk-ordered partition for clinical variables (e.g., ~1.3 percentage points for BMI at k=3). As expected, the risk-ordered partition reduced Mirai 5-year-risk heterogeneity substantially more than the embedding clusters, as it is defined by binning on predicted risk.



Figure 5. Sankey diagrams of cross-partition flow between embedding-derived clusters (left) and risk-ordered partitions (right), at k=3 and k=6. Ribbon width is proportional to the number of women shared between a cluster and a risk-ordered group. Concordance was high at k=3 (ARI = 0.860), with predominantly diagonal flow indicating the embedding clusters can be closely reproduced by ordering women on predicted risk alone; agreement was more moderate at k=6 (ARI = 0.545), reflecting greater cross-partition mixing as the partition is refined. Together these support the conclusion that stable embedding clusters largely recapitulate a scalar risk ordering rather than encoding additional phenotypic structure.

8 Conclusions/References

Clustering Mirai's exam-level embeddings recovered groups that largely correspond to a BMI–density **risk continuum** rather than qualitatively distinct phenotypes. Despite high stability (ARI [7] > 0.98) and significant differentiation across all clinical variables, comparison with risk-ordered partitions showed that these groups could be approximately reproduced by ordering women by predicted risk alone (ARI = 0.86 at k = 3). **Racial and ethnic composition varied significantly across clusters (Cramér's V = 0.225), with patterns not fully explained by BMI and density, and Asian subgroups remained persistently overrepresented in high-risk clusters after clinical adjustment.**

Stable clusters are **necessary but not sufficient** evidence for meaningful substructure; **Proposed standard:** test clusters against an ordering along the representation's primary dimension (predicted risk for DL models) to separate genuine phenotypes from partitions of a gradient.

References: [1] Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y.-L., Lamb, L., Hughes, K., Lehman, C., & Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13(578), eaba4373. <https://doi.org/10.1126/scitranslmed.aba4373>; [2] Kontos, D., Winham, S. J., Oustimov, A., Pantalone, L., Hsieh, M.-K., Gastounioti, A., Whaley, D. H., Hruska, C. B., Kerlikowske, K., Brandt, K., Conant, E. F., & Vachon, C. M. (2019). Radiomic Phenotypes of Mammographic Parenchymal Complexity: Toward Augmenting Breast Density in Breast Cancer Risk Assessment. *Radiology*, 290(1), 41–49. <https://doi.org/10.1148/radiol.2018180179>; [3] Winham, S. J., McCarthy, A. M., Scott, C. G., Gastounioti, A., Horig, H., Norman, A. D., Mankowski, W. C., Pantalone, L., Jensen, M. R., Acclavatti, R. J., Maidment, A. D. A., Cohen, E. A., Brandt, K. R., Conant, E. F., Kerlikowske, K. M., Kontos, D., & Vachon, C. M. (2025). Radiomic Parenchymal Phenotypes of Breast Texture from Mammography and Association with Risk of Breast Cancer. *Radiology*, 315(2), e240281. <https://doi.org/10.1148/radiol.240281>; [4] Xu, Y., Jha, A., Chen, Y., Shen, Y., & Heacock, L. (2026). Revealing Mammographic Phenotypes in Deep Learning Breast Cancer Risk Models: Medical Imaging with Deep Learning-Short Papers. 2026; [5] Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1), 5233. <https://doi.org/10.1038/s41598-019-41695-z>; [6] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*; [7] Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218. <https://doi.org/10.1007/bf01908075>