Data collection protocol: all datasets were developed with the same IRB-approved protocol. Cases (unilateral breast disease) were either: (1) women attending the breast clinics at Moffitt Cancer Center (MCC) diagnosed with breast cancer (type-1) or (2) attendees of surrounding area clinics sent to MCC for breast cancer treatment or diagnostic purposes and found to have breast cancer (type-2). Cases have pathology verified unilateral (first time) breast cancer. Controls were attendees of MCC with no history of breast cancer. Controls were individually matched to cases on age (± 2 years), hormone replacement therapy (HRT) usage and current duration, screening history, and mammography unit. The HRT match was based on status of current users or non-users. Nonusers included women that have not taken HRT for at least two years. If a case was a current HRT user, the control was matched on this duration (± 2 years). Controls were matched by screening history using a three-category classification. Group 1 included women with prior screening history by any means; the duration between the last screening and the study image date must be no more than 30 months. Group 2 included women with a screening history that does not fit within in Group 1 or Group 3. Group 3 included women with no screening history. We used mammograms in cranial caudal (CC) orientation as study images. The unaffected breast was used as the study image for cases (image acquired before treatment) and the matching lateral breast for controls. Women that had breast implants were excluded from this study. Cases were selected retrospectively (type-1) via electronic medical records search or recruited (type-2). Controls were selected retrospectively via electronic medical records search. Multiple suitable controls were matched to a given case and one control was selected randomly for the study.

Limitations: sampling of cases and controls was not population-based, but rather a mixture of cases ascertained at an NCI-designated comprehensive cancer center inclusive of referrals from the community. There is no evidence from our studies that the cases are not representative, but findings should be replicated in a population-based study. The data fields allow for selection of the population-based cases (discounting the referrals) but will reduce the case-control numbers. Image data from the General Electric Senographe 2000D full field digital mammography unit does not include women with large breasts due to the unit's x-ray detector design limitations. Images may contain artifacts such as nipple markers, mole markers, biopsy clips, and scar markers. These artifacts are documented in the data fields. All images were visually inspected and approved for automated processing. Here, a judgement was made to exclude a sample with too many markers or to include the sample because artifacts were deemed negligent.