All BSSA digital screening mammograms[1] Jan 2010 - Dec 2016
207 701 / 452 380 / 1 943 676
(n individuals/rounds/images)

Excluded:
- Mammogram dicoms with view other than 'CC' or 'MLO'
(eg 'LM' - n images=55 070)
- Unclear or missing patient ID
(n rounds=10)

'CC' and 'MLO' Mammograms
207 691/ 435 482 / 1 888 529
(n rounds/exams/images)

All individuals with biopsy proven lesions.
3264 / 3265 / 25 335
(n individuals/rounds/images)

Control individuals
204 427 / 418 363 / 1 863 194

Excluded individuals:
- History of biopsy (n=192)
- Breast implants (n=2 188)
- Breast symptoms (n=18 103)
- Only one round of screening (n=50 054)

Excluded rounds:
- Most recent round with no subsequent follow-up (n=130 457)

Random[3] split into training (70%), validation (15%), test (15%) subsets

Control pool
140 917 / 211 462 / 933 727
(n individuals/rounds/images)

Train cases

Train control pool

Validation cases

Val. control pool

Test cases

Test control pool

Age-matched

Age-matched

Age-matched

Excluded: rounds with less than 4 images (left and right MLO and CC) per individual per round

Training set (n=4478)
- individuals with malignant biopsy (n=1978)
- Malignancy prevalence: 44.17%
- individuals with benign biopsy (n=238)
- Age-matched controls (n=2262)
- Total images: 20 193

Validation set (n=963)
- individuals with malignant biopsy (n=434)
- Malignancy prevalence: 45.07%
- individuals with benign biopsy (n=41)
- Age-matched controls (n=488)
- Total images: 4310

Test sets
- individuals with malignant biopsy (n=425)
- individuals with benign biopsy (n=44)

Test set 1 - balanced (n=959)
- Age-matched controls (n=490)
- Malignancy prevalence: 44.31%
- Total images: 4229

Test set 2 - Approx NYU incidence (n=3254)
- Age-matched controls (n=2785)
- Malignancy prevalence: 13.06%
- Total images: 14 415