Codebook for CSAW-CC anonymous dataset

**Copyright holder and curators**

This dataset was curated by principal investigator Fredrik Strand (fredrik.strand@ki.se), MD PhD, at the Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden. Radiologist annotations were performed by a breast radiologist at Capio S:t Göran Hospital and Karolinska Institutet, using in-house software developed by Kevin Smith, MSc PhD, associate professor, at SciLifeLab/KTH.

**Ethics**

The ethical review board of Stockholm allowed the research and waived the requirement for individual informed consent (EPN 2016/2600-31) with additions approved by the Ethical Review Authority of Sweden (EPM 2019-01946, EPM 2019-03638, EPM 2021-01030).

**Data sources**

Data from Karolinska University Hospital is the main information source. In addition, data on cancer characteristics that originate from Karolinska University Hospital was collected through the Regional Cancer Center Stockholm-Gotland. Data has been further process by the copyright holders and curators above.

**Conditions for use**

**General conditions:** You are allowed to download and use the data for research purposes free of charge after requesting the data.
**Specific conditions:** The ICMJE criteria regarding inviting the data curators as co-authors of publications must be followed.

**CC BY:** The dataset is published under a CC BY license. \*

**Anonymization**

Anonymization of clinical data was achieved by redaction of any personal identifiers, and by redaction of all but the most essential parameters, by perturbation of data and by binning of numerical data. Through this practice a k-anonymity of 10 was achieved. No key to re-connect the anonymous ID numbers to the original person-identifiers have been retained and therefore re-matching is not possible. The mammographic images have been assessed not to allow identification of individuals based on several observations: a mammogram is an imprecise image of the body due to projection of all tissue to a 2-dimensional matrix and cannot be used to identify persons by visual observation, mammograms of the same person changes over time, the body part examined is not a body part for which optical images are commonly available, even for a trained breast radiologist it is not possible to use a mammogram to identify a person that they know from everyday life. A visual inspection of included mammograms do not reveal any additional information that can be used to identify individuals.

**Selection criteria**

*Cases of breast cancer*: All first-time breast cancer diagnoses from Karolinska University Hospital (Stockholm, Sweden) uptake area between late 2008 and Dec 31, 2015.

*Healthy controls*: Random selection of healthy controls (Stockholm Sweden) examined between May 6, 2008 and Dec 23, 2016.

*Not included:* Women outside the screening age range 40-74. Women who did not attend mammography screening. Examinations that did not include all four standard views. A random selection of thirty percent of the population is held back and not shared (can be used for validation in collaboration).

**Images**

**Mammograms (DICOM format)**: All screening mammograms between the above dates for all included individuals were full-field digital mammograms acquired on Hologic equipment. For women who were diagnosed, images after the date of diagnosis are not included. The filename structure contains the following information: [anon\_patientid] + ’\_20990909\_’ + [laterality, R=right, L=left] + ’\_’ + [projection: MLO or CC] + [random number].dcm’.

**Annotations (PNG format)**: A breast radiologist (KD) annotated all mammograms with visible cancer. The most recent prior screening mammograms (if available) were also annotated – if there were cancer signs those were annotated, if no cancer signs were present in the prior images a dot was placed to indicate the corresponding location where the cancer later arose.

**Variable grouping**

Variables with “x\_” prefix are demographic or cancer data which are reported on a patient level, and recur on all rows pertaining to a specific patient regardless of timing of the exam (before, or at diagnosis).

Variables with “rad\_” prefix are radiological data which are reported on exam level, and therefore recur on the four rows pertaining to a specific patient and mammography examination date.

Variables imagelaterality and viewposition are reported on image (file name) level.

**List of variables in csv file *CSAW-CC\_breast\_cancer\_screening\_data.csv***

anon\_patientid = anonymous patient ID

exam\_year = year of the mammography exam

anon\_filename = mammogram image file name

x\_age

1 = 40-55 years of age at mammography

2 = 55+ years of age at mammography

x\_case

0 = no cancer during study time
1 = cancer during study time (see rad\_timing below)

x\_cancer\_laterality

left = cancer diagnosed in left breast

right = cancer diagnosed in right breast

x\_type

1 = in situ only

2 = invasive <=15 mm

3 = invasive >15 mm

x\_lymphnode\_met

0 = no

1 = yes

rad\_timing = timing of radiology exam

1 = <60 days from screening to diagnosis (screen-detected cancer)

2 = 60-729 days (interval cancer)

3 = 730+ days (priors/other)

rad\_r1 = assessment by radiologist 1

0 = healthy

1 = discuss

rad\_r2 = assessment by radiologist 2

0 = healthy

1 = discuss

rad\_recall = decision to recall for further work-up

0 = normal

1 = recall

rad\_recall\_type\_right = for the right breast, reason for recall (patient symptom or mammographic finding)

1 = radiological

2 = symptomatic

rad\_recall\_type\_left = for the left breast, reason for recall (patient symptom or mammographic finding)

1 = radiological

2 = symptomatic

imagelaterality

left = image of left breast

right = image of right breast

viewposition

cc = craniocaudal projection

mlo = mediolateral oblique projection

libra\_breastarea

Breast area in cm2 estimated by the Libra software (version 1.0.4, University of Pennsylvania, Philadelphia, Pa)

libra\_densearea

Dense area in cm2 estimated by the Libra software (version 1.0.4, University of Pennsylvania, Philadelphia, Pa)

libra\_percentdensity

Percent density estimated by the Libra software (version 1.0.4, University of Pennsylvania, Philadelphia, Pa)

\*Before 2023-08-25 the dataset was published under a CC BY-NC-SA license.