CSAW-CC (mammography) – a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer

SND-ID: 2021-204-1. **Version**: 1. **DOI**: https://doi.org/10.5878/45vm-t798

Associated documentation

CSAW-CC_breast_cancer_screening_data.csv *(8.78 MB)*

CSAW-CC_Readme_annon_230516.docx *(28.37 KB)*

Citation

Strand, F. (2022) CSAW-CC (mammography) – a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer (Version 1) [Data set]. Karolinska Institutet. Available at: https://doi.org/10.5878/45vm-t798

Alternative title

Cohort of Screen-age Women - Case control (CSAW-CC)

Creator/Principal investigator(s)

Fredrik Strand - Karolinska Institutet, Department of Oncology-Pathology

Research principal

Karolinska Institutet - Department of Oncology-Pathology

Principal's reference number

4-3790/2016

Description

The dataset contains x-ray images, mammography, from breast cancer screening at the Karolinska University Hospital, Stockholm, Sweden, collected by principal investigator Fredrik Strand at Karolinska Institutet. The purpose for compiling the dataset was to perform AI research to improve screening, diagnostics and prognostics of breast cancer.

The dataset is based on a selection of cases with and without a breast cancer diagnosis, taken from a more comprehensive source dataset.

1,103 cases of first-time breast cancer for women in the screening age range (40-74 years) during the included time period (November 2008 to December 2015) were included. Of these, a random selection of 873 cases have been included in the published dataset.

A random selection of 10,000 healthy controls during the same time period were included. Of these, a random selection of 7,850 cases have been included in the published dataset.

For each individual all screening mammograms, also repeated over time, were included; as well as the date of screening and the age. In addition, there are pixel-level annotations of the tumors created by a breast radiologist (small lesions such as micro-calcifications have been annotated as an area).

Annotations were also drawn in mammograms prior to diagnosis; if these contain a single pixel it means no cancer was seen but the estimated location of the center of the future cancer was shown by a single pixel annotation.

In addition to images, the dataset also contains cancer data created at the Karolinska University Hospital and extracted through the Regional Cancer Center Stockholm-Gotland. This data contains information about the time of diagnosis and cancer characteristics including tumor size, histology and lymph node metastasis.

The precision of non-image data was decreased, through categorisation and jittering, to ensure that no single individual can be identified.

The following types of files are available:
- CSV: The following data is included (if applicable): cancer/no cancer (meaning breast cancer during 2008 to 2015), age group at screening, days from image to diagnosis (if any), cancer histology, cancer size group, ipsilateral axillary lymph node metastasis. There is one csv file for the entire dataset, with one row per image. Any information about cancer diagnosis is repeated for all rows for an individual who was diagnosed (i.e., it is also included in rows before diagnosis). For each exam date there is the assessment by radiologist 1, radiologist 2 and the consensus decision.
- DICOM: Mammograms. For each screening, four images for the standard views were acuqired: left and right, mediolateral oblique and craniocaudal. There should be four files per examination date.
- PNG: Cancer annotations. For each DICOM image containing a visible tumor.

Access:
The dataset is available upon request due to the size of the material. The image files in DICOM and PNG format comprises approximately 2.5 TB.
Access to the CSV file including parametric data is possible via download as associated documentation.

Data contains personal data

No

Language

English

Unit of analysis

Individual/Patient

Population

Women 40-74 years of age who were invited to mammography screening

Study design

Case-control study

Description of study design

Case-control cohort regarding breast cancer diagnosis. All 1103 cases of first-time breast cancer for women in the screening age range (40-74 years) during the included time period (late 2008 to Dec 31, 2015) were included. A random selection of 10,000 healthy controls during the same time period

were included. Of these, a random selection of 873 of diagnosed cases and of 7850 healthy controls, designated “non-hidden”, have been included in the published dataset.

Sampling procedure

Total universe/Complete enumeration
Probability: Systematic random
Cases: Consecutive breast cancer diagnoses within the population of women who were invited to mammography screening before Dec 31, 2015.
Controls: Randomly selected women who were not diagnosed with breast cancer before Dec 31, 2015.

Time period(s) investigated

2008 – 2015

Variables

19

Number of individuals/objects

8723

Data format / data structure

Numeric

Text

Still image

Data collection 1

- Mode of collection: Registry extract and/or access to biobank sample
- Time period(s) for data collection: 2008 – 2015
- Data collector: Karolinska University Hospital
- Source of the data: Registers/Records/Accounts: Medical/Clinical, Registers/Records/Accounts

Data collection 2

- Mode of collection: Registry extract and/or access to biobank sample
- Time period(s) for data collection: 2008 – 2015
- Data collector: Regional Cancer Centre Stockholm-Gotland
- Source of the data: Registers/Records/Accounts: Medical/Clinical, Registers/Records/Accounts

Geographic spread

Geographic location: Stockholm County

Geographic description: The geographical uptake area of breast cancer screening at the Karolinska University Hospital in Stockholm, Sweden

Lowest geographic unit

Region

Highest geographic unit

Region

Responsible department/unit

Department of Oncology-Pathology

Contributor(s)

Kevin Smith - Royal Institute of Technology, SciLifeLab

Karolinska University Hospital, Breast Radiology

SciLifeLab - Royal Institute of Technology

Regional Cancer Centre Stockholm-Gotland

Ethics Review

Stockholm - Ref. 2016/2600-31

Swedish Ethical Review Authority - Ref. 2021-01030

Swedish Ethical Review Authority - Ref. 2019-03638

Swedish Ethical Review Authority - Ref. 2019-01946

Research area

Cancer and oncology (Standard för svensk indelning av forskningsämnen 2011)

Radiology, nuclear medicine and medical imaging (Standard för svensk indelning av forskningsämnen 2011)

Keywords

Breast neoplasms, Mammography

Publications

Dembrower, K., Liu, Y., Azizpour, H., Eklund, M., Smith, K., Lindholm, P., & Strand, F. (2020). Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. Radiology, 294(2), 265–272. https://doi.org/10.1148/radiol.2019190872
DOI: https://doi.org/10.1148/radiol.2019190872
URN: urn:nbn:se:kth:diva-267834

Dembrower K, Lindholm P, Strand F. A Multi-million Mammography Image Dataset and Population-Based Screening Cohort for the Training and Evaluation of Deep Neural Networks-the Cohort of Screen-Aged Women (CSAW). J Digit Imaging. 2019.
DOI: https://doi.org/10.1007/s10278-019-00278-0

Dembrower, K., Wahlin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., & Strand, F. (2020). Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload : a retrospective simulation study. The Lancet Digital Health, 2(9), E468–E474. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-281510
DOI: https://doi.org/10.1016/S2589-7500(20)30185-0
URN: urn:nbn:se:kth:diva-281510

Salim, M., Wåhlin, E., Dembrower, K., Azavedo, E., Foukakis, T., Liu, Y., Smith, K., Eklund, M., & Strand, F. (2020). External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncology, 6(10), 1581. https://doi.org/10.1001/jamaoncol.2020.3321

DOI: https://doi.org/10.1001/jamaoncol.2020.3321
URN: urn:nbn:se:kth:diva-284972

Accessibility level

Access to data through SND
Access to data is restricted

Use of data

Things to consider when using data shared through SND

License

CC BY 4.0

Versions

Version 1. 2022-04-22

Homepage

Research Group of Principal Investigator

Contact for questions about the data

Data Access Unit
rdo@ki.se

Download metadata

DataCite
DDI 2.5
DDI 3.3
DCAT-AP-SE 2.0
JSON-LD
PDF
Citation (CLS)

Published: 2022-04-22
Last updated: 2023-12-01