MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] The main goal in creating this dataset was to enable the development of models capable of identifying
screening participants in mammography screening whose mammograms cannot easily be assessed due to a high level of mammographic masking, a phenomenon that occurs when potential cancer could largely be obscured by the surrounding tissue in the breast. As a result, breast cancer in these participants is more likely to be missed during regular mammography. More sensitive imaging technologies such as MRI are too costly to be provided for all participants visiting a clinic. Due to the large number of mammography images that are taken at clinics, there exists a need to develop an AI model that could help identifying screening participants in higher needs of MRI. To develop such an AI model, we noticed a lack of mammographic images containing assessment of masking level directly made by expert radiologists. Although public mammographic datasets exist, none of them exactly contains direct potential of masking in mammograms assessed by radiologists. Our aim was to fill the gap by collecting a dataset that merely focuses on mammographic masking. CSAW-M helps to automate identifying of low- and high-masking mammograms

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] The dataset was created in a joint collaboration of researchers from KTH Royal Institute of Technology, Karolinska Institutet, Karolinska University Hospital, and S:t Görans Hospital in Stockholm.

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] This work was partially supported by MedTechLabs https://www.medtechlabs.se/, the Swedish Research Council (VR) 2017-04609, and Region Stockholm HMT 20200958.

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] The dataset comprises mammographic images, together with metadata that is provided as CSV files. The metadata includes masking potential labels collected from five experts, image acquisition parameters, clinical endpoints i.e. cancer attributes and density measures. More details can be found in the main paper.

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] There are 10,020 screening participants in total, and each participant has 1 mammogram from the MLO view of the breast. 9,523 of the images are in CSAW-M training set with one annotation per image, while the rest 497 images are from a public test set where each image has annotations from 5 experts

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

[answer7] CSAW-M is a subset of CSAW, a large population-level cohort of screening mammograms [25]. We exclude images that are: a) from patients with implants; b) biopsy images; c) mammograms that are not vendor post-processed; d) mammograms with aborted exposure; e) mammograms not taken by X-ray photoconductor; f) earlier mammograms when there are duplicates in the same exam. We sample screening participants with complete mammography exams taken in Karolinska University Hospital and with Hologic manufacturer. We sampled from the participants according to the procedure mentioned in Section 2 in a way that more mammograms with extreme density values are included (which are more clinically interesting), so the sampled mammograms are not necessarily representative of the larger set. This was done because mammograms in the tails of the percent density distribution (very dense or very fatty) are of the highest clinical interest.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Images in PNG format, certain DICOM acquisition attributes that can be used for preprocessing, clinical endpoints, and masking annotations from 5 experts (as detailed in Table 2). Training images have one annotation while test images have 5 annotations per image.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Yes, the labels are masking levels (from 1-8) of each instance annotated by 5 experts, together with certain clinical endpoints, i.e. interval or large invasive cancer.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] No, there is no missing information. The information is complete for all individual instances.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] Yes, the annotations are explicitly applied to the images, which were shown directly to the experts.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] Yes, we have recommended training and testing splits. We have benchmarked mammographic masking of cancer on suggested testing splits where there are 5 annotations per image (the median was chosen as ground truth), as opposed to the training set that contains 1 annotation per image. There is no recommended development/validation split. However, we have provided the cross-validation folds that we used when developing the models.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] Yes. Experts may have made wrong button clicks or errors in their comparisons. Similarly, there may
be clerical errors matching patients with their clinical endpoints.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] The dataset is hosted by the https://scilifelab.figshare.com/. It has restricted access where users must submit their request, after which the access to the actual files could be granted. See Appendix K for more details.

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15] Yes, the dataset contains information related to the health status of individuals. The information has
been reduced in order not to allow the identification of any individual. In our assessment, the dataset contains only de-identified information.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] No. Our dataset mainly contains mammography images, and there is nothing offensive, insulting, or
threatening.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] In Sweden, individuals with female personal identity numbers are invited for mammographic screening. Our dataset contains mammogramgraphic screenings from screening participants 40 to 74 years of age. Racial information is not collected in Sweden.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] No, we have taken appropriate measures to ensure it is not possible. The measures include: (1) we removed all individual identifiers from the data, (2) we down-sampled the mammograms, (3) we removed all unnecessary acquisition attributes –DICOM headers–, (4) we simplified the continuous

tumor size attribute to a binary outcome, and (5) we anticipated a gated release mechanism to approve users based on their information and project goals before granting access to the data. Users are also required to explicitly agree not to attempt to de-identify any individuals from the dataset.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

[answer19] The cancer images in our dataset are accompanied with the clinical outcome of the screening corresponding to that image, i.e. whether they are diagnosed with interval or large invasive cancer. These attributes, however, are available in our dataset in a binary form and the screening participants are de-identified.

[question20] Any other comments?

[answer20] The source dataset is extracted from a large cohort containing millions of mammograms, collected every 18 to 24 months from screening participants aged 40 to 74 in Stockholm county area. The dataset contains around 10,020 mammograms taken from 10,020 participants.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] The source images of our dataset are in DICOM format with DICOM metadata. Each patient is linked to the Regional Cancer Registry to define clinical endpoints such as whether a screening participant was healthy or had been diagnosed with breast cancer. Mammograms were shown to five experts to assign masking annotations.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] We designed a user interface through which we showed two images alongside each other and asked
experts to do pair-wise comparisons and select the image that is harder to assess. It was validated to be bug-free by running several tests with experts before the collection process began.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] The data is sampled from CSAW as explained in Section 2. The sampling was done in way to include more mammograms with very low or very high percent density measure as these are the most clinically interesting images.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] Researchers from KTH Royal Institute of Technology, Karolinska Institutet, Karolinska University Hospital, and S:t Görans Hospital in Stockholm were involved in the data collection. All participants were compensated for their time in the course of their normal research activities.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] The mammograms were collected during regular mammography screening between 2008 and 2015 at Karolinska University Hospital. The creation of the CSAW-M dataset, including developing the annotation tool, receiving annotations from experts, cleaning data etc. was initiated in June 2020 and lasted until November 2020.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26] The Regional Ethical Review Board in Stockholm has approved the research. Also, a dedicated agreement between Karolinska Institutet and KTH Royal Institute of Technology has been made to publish the data.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] NO

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

[answer27] The source mammograms were collected by Karolinska University Hospital, the clinical labels were collected by the Regional Cancer Center, and masking labels were collected by showing mammograms to five experts for annotation.

[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

[answer28] No. The need for informed consent was waived by the Ethical Review Board.

[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

[answer29] No. The need for informed consent was waived by the Ethical Review Board.

[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so,

please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30] Not applicable.
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31] We consulted an expert in GDPR and dealing with personal data from Karolinska Institutet, and our concerns regarding privacy were cleared.
[question32] Any other comments?
[answer32] No

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] Data preprocessing was done in our baseline implementations. The source images of our dataset were DICOM files whose pixel values we saved as raw PNG images. Using DICOM metadata, we did preprocessing to generate PNG images. Please refer to Section 2 for more details about image preprocessing.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] The "raw" data was saved as PNG, and we also provide the preprocessing script that was used in our baseline implementation for reproducibility.
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
[answer35] We used Python standard libraries and the preprocessing script is available in the Github repo of the project: https://github.com/yueliukth/CSAW-M/.
[question36] Any other comments?
[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.
[answer37] Yes. The masking model, together with two other models that we developed to perform breast cancer

risk prediction and cancer detection, are combined into a single comprehensive model. This clinical workflow is currently implemented at Karolinska University Hospital in a clinical study to help identify screening participants who are most likely to benefit from additional MRI screening.

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38] No.

[question39] What (other) tasks could the dataset be used for?

[answer39] First, our dataset has annotations that are ordinally related and can be used to study ordinal classification or point-wise ranking tasks. Specifically, our public test set contains 5 annotaitons per image, which make it a useful resource to study human noise and bias. Moreover, our dataset which contains more than 10,000 mammograms, is significantly larger than other public mammography datasets (see Table 1 in the main paper). It can be used for pretraining deep learning models that would be used in other downstream tasks in a similar domain to mammography images (for more effective transfer learning). And last but certainly not least, we included clinical endpoints as our metadata, making it valuable in clinical studies. We have shown in the paper that our ResNet-34 models trained on estimating masking potential perform better than the breast density counterparts in identifying screening participants diagnosed with interval and large invasive cancers, without being explicitly trained for these tasks. This shows a great promise for the usefulness of our collected labels and motivates developing better models for estimating masking level.

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] We are aware of the fact that biases exist inherently in our data collection, for the following reasons:
(a) the data was extracted from a certain population, period and region, with certain manufacturers, (b) the annotations were made by radiologists from a certain region, (c) we randomly sampled screening participants and intentionally selected breasts that are denser or fattier which resulted in a distribution that is not representative of the real population. We note that clinical studies are crucially required before deploying models in any clinical processes.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] In the main article, we have noted that our dataset is not aimed for developing/evaluating cancer detection models, as the cancer images in CSAW-M are chosen to be contralateral to cancer laterality, i.e. the breast that does not contain tumor was selected (please refer to Section 2 for motivation).

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

[answer43] SciLifeLab Data Repository, who hosts our dataset, is currently relying on the Figshare service, but plans to move data to its own storage servers soon. This does not change availability of the dataset in any way, nor does it impose additional restrictions by any third party. SciLifeLab Data Repository is affiliated with KTH Royal Institute of Technology and the hosting of our dataset is guaranteed there.

[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

[answer44] The dataset webpage could be found with this DOI 10.17044/scilifelab.14687271. All the instructions on how to access the data is clearly mentioned on the dataset landing page, which contains the actual data files along with metadata to help users better understand how to use the data.

[question45] When will the dataset be distributed?

[answer45] The dataset has already been distributed with this DOI 10.17044/scilifelab.14687271.

[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

[answer46] Yes. Please visit the dataset home page for details about the license and terms.

[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

[answer47] There is no third parties imposed IP-based or other restrictions on the data associated with the instances.

[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

[answer48] There are no export controls or other regulatory restrictions on this dataset to the best of our knowledge.

[question49] Any other comments?

[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?

[answer50] The data is currently supported/hosted by https://scilifelab.figshare.com/ (the support letter could be seen on the final page of this article). The infrastructure for hosting and maintaining the data is guaranteed to be supported by the repository.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?

[answer51] The owners of the dataset could be contacted through either of the following email addresses: yue3@kth.se and sorkhei@kth.se.

[question52] Is there an erratum? If so, please provide a link or other access point.

[answer52] There is no erratum for the dataset.

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

[answer53] At the moment, there is no plan for any updates. In case the dataset is updated, the most recent version of it could be seen on the dataset website (previous versions will still be visible), and the DOI will also change accordingly with respect to the version.

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

[answer54] In case of retention, the data will be deleted and a new version that addresses the issue will be re-uploaded, in which case we ask users to delete their old copy of data (our ToU covers this). This is communicated clearly to the users.

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

[answer55] If there is a change in the version of the dataset, previous versions will still be hosted and supported on the website. We will announce the change of version as explicit as possible on the website.

[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] We always welcome if experts in the area of mammography are interested in contributing to our dataset by assessing mammographic potential masking of tumor in our mammograms. Code for our annotation tool with complete instructions on how to use it is publicly available at https://github.com/MoeinSorkhei/CSAW-M_Annotation_Tool/. For further discussion, experts are very welcome to contact us using the contact info on the website. We will then compare the received annotations against our ground truth using the same metrics we used in the paper. Finally, the annotations and the comparison against our ground-truth will be made publicly available on our website, acknowledging the contribution. We are also interested in receiving BI-RADS annotations. We would be happy to discuss any other possible contributions not mentioned here. We note, however, that although contributions will be made publicly visible on our website, they do not result in any change in the authors of the dataset.

[question57] Any other comments?

[answer57] No.