# A DICOM dataset for evaluation of medical image de-identification (Pseudo-PHI-DICOM-Data)

# Summary

Open access or shared research data must comply with (HIPAA) patient privacy regulations. These regulations require the de-identification of datasets before they can be placed in the public domain. The process of image de-identification is time consuming, requires significant human resources, and is prone to human error. Automated image de-identification algorithms have been developed but the research community requires some method of evaluation before such tools can be widely accepted. This evaluation requires a robust dataset that can be used as part of an evaluation process for de-identification algorithms.

We developed a DICOM dataset that can be used to evaluate the performance of de-identification algorithms. DICOM image information objects were selected from datasets published in TCIA. Synthetic Protected Health Information (PHI) was generated and inserted into selected DICOM data elements to mimic typical clinical imaging exams. The evaluation dataset was de-identified by a TCIA curation team using standard TCIA tools and procedures. We are publishing the evaluation dataset (containing synthetic PHI) and de-identified evaluation dataset (result of TCIA curation) in advance of a potential competition, sponsored by the National Cancer Institute (NCI), for de-identification algorithm evaluation, and de-identification of medical image datasets. The evaluation dataset published here is a subset of a larger evaluation dataset that was created under contract for the National Cancer Institute. This subset is being published to allow researchers to test their de-identification algorithms and promote standardized procedures for validating automated de-identification.

**Acknowledgements**

We would like to acknowledge the National Cancer Institute for funding and actively participating in the project that generated the evaluation datasets being published here and the TCIA curation team, led by Ms. Geri Blake, who curated this data. Original data came from multiple institutions and multiple TCIA image collections.

**Data Access**

## Data Access

| Data Type | Download all or Query/Filter | License |
|---|---|---|
| Images, (DICOM, 609 MB)<br>Evaluation dataset | *Download* *Search*<br>(Download requires the NBIA Data Retriever) | CC BY 4.0 |
| Images, (DICOM, 606 MB)<br>De-identified Evaluation dataset | *Download* *Search*<br>(Download requires the NBIA Data Retriever) | CC BY 4.0 |
| Patient Mapping (csv, 0.6 kB)<br>Evaluation/De-identified | *Download* | CC BY 4.0 |
| UID Mapping (csv, 213 kB)<br>Evaluation/De-identified | *Download* | CC BY 4.0 |

Click the Versions tab for more info about data releases.

Please contact help@cancerimagingarchive.net with any questions regarding usage.

## Additional Resources for this Dataset

The NCI Cancer Research Data Commons (CRDC) provides access to additional data and a cloud-based data science infrastructure that connects data sets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data.

- Imaging Data Commons (IDC) (Imaging Data)

**Detailed Description**

## Detailed Description

| Image Statistics | |
|---|---|
| Modalities | CR, CT, DX, MG, MR, PT |
| Number of Patients | 42 |
| Number of Studies | 44 |
| Number of Series | 52 |
| Number of Images | 3386 |
| Images Size (GB) | 1.2 |

There are 21 patients, 22 studies, 26 series but the patient ids, study instance uids, and series instance uids are different between the 2 datasets thus resulting in a double count.

**Citations & Data Usage Policy**

## Citations & Data Usage Policy

Users must abide by the TCIA Data Usage Policy and Restrictions. Attribution should include references to the following citations:

**Data Citation**

Rutherford, M., Mun, S.K., Levine, B., Bennett, W.C., Smith, K., Farmer, P., Jarosz, J., Wagner, U., Farahani, K., Prior, F. (2021). A DICOM dataset for evaluation of medical image de-identification (Pseudo-PHI-DICOM-Data) [Data set]. The Cancer Imaging Archive. DOI: https://doi.org/10.7937/s17z-r072

**Publication Citation**

Rutherford, M., Mun, S.K., Levine, B., Bennett, W.C., Smith, K., Farmer, P., Jarosz, J., Wagner, U., Freyman, J., Blake, G., Tarbox, L., Farahani, K., Prior, F. (2021). A DICOM dataset for evaluation of medical image de-identification, Nature Scientific Data. DOI: 10.1038/s41597-021-00967-y.

**TCIA Citation**

Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, December, 2013, pp 1045-1057. DOI: 10.1007/s10278-013-9622-7

## Other Publications Using This Data

TCIA maintains a list of publications which leverage TCIA data. If you have a manuscript you'd like to add please contact the TCIA Helpdesk.

**Versions**

## Version 2 (Current): Updated 2021/04/07

| Data Type | Download all or Query/Filter |
|---|---|
| Images, (DICOM, 609 MB)<br>Evaluation dataset | *Download* *Search*<br>(Download requires the NBIA Data Retriever) |
| Images, (DICOM, 606 MB)<br>De-identified Evaluation dataset | *Download* *Search*<br>(Download requires the NBIA Data Retriever) |
| Patient Mapping (csv)<br>Evaluation/De-identified | *Download* |
| UID Mapping (csv)<br>Evaluation/De-identified | *Download* |

Note: Removed head imaging from 8 series.

## Version 1: Updated 2021/01/31

| Data Type | Download all or Query/Filter |
|---|---|
| Images, (DICOM, 653 MB)<br>Evaluation dataset | *Search*<br>(Download requires the NBIA Data Retriever) |
| Images, (DICOM, 648 MB)<br>De-identified Evaluation dataset | *Search*<br>(Download requires the NBIA Data Retriever) |
| Patient Mapping (csv)<br>Evaluation/De-identified | *Download* |
| UID Mapping (csv)<br>Evaluation/De-identified | *Download* |