

# MosMedData: Chest CT Scans with COVID-19 Related Findings

---



РАДИОЛОГИЯ МОСКВЫ  
ДИАГНОСТИКА БУДУЩЕГО

This dataset contains anonymised human lung computed tomography (CT) scans with COVID-19 related findings, as well as without such findings. A small subset of studies has been annotated with binary pixel masks depicting regions of interests (ground-glass opacifications and consolidations). CT scans were obtained between 1st of March, 2020 and 25th of April, 2020, and provided by medical hospitals in Moscow, Russia.

## DISCLAIMER

This dataset is intended to be used as:

- educational material for medical imaging specialists showing intrinsic radiological signs of COVID-19 infection;
- a dataset for development, training and testing of AI-based services that ;
- information source for medical specialists and broad audience.

You are free to **share** this dataset, which means you can copy and redistribute the material in any medium or format, **under the following terms:**

- you must give appropriate credit, such as:
  - authors;
  - their affiliations;
  - copyright;
  - permanent link to the dataset.
- you must share a licence or provide a link to it.

You **must not:**

- use this dataset for commercial purposes;
- distribute the modified material if you remix, transform, or build upon the dataset;
- apply legal terms or technological measures that legally restrict others from doing anything the license permits.



## General Information

---

### Dataset Title

MosMedData: Chest CT Scans with COVID-19 Related Findings

### Internal Code

COVID19\_1110

## Annotation Class

2-C, 2-A

## Keywords

computed tomography, CT, pulmonary, viral, infection, lungs, chest, COVID-19

## Language

English, Russian

## Funding Sources

Internal funding

## Dataset Version

1.0

## Permanent link

[https://mosmed.ai/datasets/covid19\\_1110](https://mosmed.ai/datasets/covid19_1110)

## Release Date

28.04.2020

## Affiliation and Contributors

---

### Contributors

- Morozov, Sergey<sup>1</sup>
- Andreychenko, Anna<sup>1</sup>
- Blokhin, Ivan<sup>1</sup>
- Vladzimirsky, Anton<sup>1</sup>
- Gelezhe, Pavel<sup>1</sup>
- Gombolevskiy, Victor<sup>1</sup>
- Gonchar, Anna<sup>1</sup>
- Ledikhova, Natalya<sup>1</sup>
- Pavlov, Nikolay<sup>1</sup> ([n.pavlov@npcmr.ru](mailto:n.pavlov@npcmr.ru))
- Chernina, Valeriya<sup>1</sup>

### Affiliation

1. Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department.

## Data Structure

```
.
|-- dataset_registry.xlsx
|-- LICENSE
|-- README_EN.md
|-- README_RU.md
|-- README_EN.pdf
|-- README_RU.pdf
|-- masks
|   |-- study_BBBB_mask.nii.gz
|   |-- ...
|   `-- study_BBBB_mask.nii.gz
`-- studies
    |-- CT-0
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-1
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-2
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-3
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    `-- CT-4
        |-- study_BBBB.nii.gz
        |-- ...
        `-- study_BBBB.nii.gz
```

- `README_EN.md` and `README_RU.md` contain general information about the dataset; they have been saved in `Markdown` format in English and Russian languages, respectively. `README_EN.pdf` and `README_RU.pdf` contain the same information but have been saved in `PDF` format for the ease of convenience.
- `LICENSE` file contains full description of Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License
- `dataset_registry.xlsx` is a spreadsheet with full list of studies included in the dataset as well as relative paths to a study file and to a binary mask, if present.
- `studies` directory contains directories named as `CT-0`, `CT-1`, `CT-2`, `CT-3`, and `CT-4` (for more information see below). Each directory contains studies in `NIfTI` format, that have been saved in `Gzip` archive. Each study has a unique name like `study_BBBB.nii.gz`, where `BBBB` is a sequential number of the study in the whole dataset.
- `masks` directory contains binary pixel masks in `NIfTI` format, that have been saved in `Gzip` archive. Each study has a unique name like `study_BBBB_mask.nii.gz`, where `BBBB` is a number of the corresponding study.

## Data Overview

Property	Value
Number of studies, pcs.	1110
Number of patients, ppl.	1110
Distribution by sex, % (M/ F/ O)	42/ 56/ 2
Distribution by age, years (min./ median/ max.)	18/ 47/ 97

Property	Value
Number of binary pixel masks (Class A Annotation), pcs.	50
Number of studies in each category (Class C Annotation), psc. (CT-0/ CT-1/ CT-2/ CT-3/ CT-4)	254/ 684/ 125/ 45/ 2

## Data Preprocessing

- Each study corresponds to unique patient.
- Each study is represented by one series of images reconstructed into soft tissue mediastinal window.

```
SeriesDescription LIKE '%BODY%'
```

- During the DICOM -to- NIfTI conversion process only every 10th image (Instance) was preserved.

```
InstanceNumber % 10 = 0
```

## Class C Annotation Principles

Studies are distributed into [5 categories](#)<sup>1</sup>:

- **CT-0** (/studies/CT-0 directory): normal lung tissue, no CT-signs of viral pneumonia.
- **CT-1** (/studies/CT-1 directory): several ground-glass opacifications, involvement of lung parenchyma is less than 25%.
- **CT-2** (/studies/CT-2 directory): ground-glass opacifications, involvement of lung parenchyma is between 25 and 50%.
- **CT-3** (/studies/CT-3 directory): ground-glass opacifications and regions of consolidation, involvement of lung parenchyma is between 50 and 75%.
- **CT-4** (/studies/CT-4 directory): diffuse ground-glass opacifications and consolidation as well as reticular changes in lungs. Involvement of lung parenchyma exceeds 75%.

**Please note:** this distribution has been made *before* DICOM -to- NIfTI conversion and *before* selection of every 10th image (Instance).

**Please note:** this distribution has been made based on *radiologic findings only*, neither on polymerase chain reaction (PCR) test results or clinical verification.

1. (In Russian only) Лучевая диагностика коронавирусной болезни (COVID-19): организация, методология, интерпретация результатов : препринт № ЦДТ – 2020 – II. Версия 2 от 17.04.2020 / сост. С. П. Морозов, Д. Н. Проценко, С. В. Сметанина [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. 65. – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», 2020. – 78 с.

## Class A Annotation Principles

A small subset of studies (50 pcs.) have been annotated by the experts of Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. During the annotation for every given image ground-glass opacifications and regions of consolidation were selected as positive (white) pixels on the corresponding binary pixel mask. The resulting masks have been saved in NIfTI format and then transformed into Gzip archive.

The [MedSeg](#) software has been used for annotation purposes (© 2020 Artificial Intelligence AS).

## Sharing and Access Information

## License

Copyright © 2020 Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department.

This dataset is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License. See [LICENSE](#) file or follow the [link](#) for more information.

## Citation

Recommended citation:

Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzmyrskyy, A.V., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.V. and Chernina, V.Y., 2020. MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset. *arXiv preprint arXiv:2005.06465*.

## Distribution

This dataset should not be distributed without proper attribution:

- authors;
- their affiliations;
- copyright;
- permanent link to the dataset;
- license file or link to the license text.