GOVERNMENT OF MOSCOW

MOSCOW HEALTH CARE DEPARTMENT

RESEARCH AND PRACTICAL CLINICAL CENTER FOR DIAGNOSTICS AND TELEMEDICINE TECHNOLOGIES, MOSCOW HEALTH CARE DEPARTMENT

ISSN 2618-7124

BEST PRACTICES IN MEDICAL IMAGING

DATASET PREPARATION REGULATIONS AND APPROACHES TO REPRESENTATIVE DATA SAMPLING

Moscow
2022

MOSCOW CENTER
FOR DIAGNOSTICS & TELEMEDICINE

GOVERNMENT OF MOSCOW
MOSCOW HEALTH CARE DEPARTMENT

APPROVED by

the Chief Regional Radiology
and Instrumental Diagnostics
Officer of the Moscow
Health Care Department

__________ S.P. Morozov

«29» октября 2021

RECOMMENDED by

the Research Expert Council
of the Moscow Health Care
Department No. 1

«21» января 2021
2022

DATASET PREPARATION REGULATIONS AND APPROACHES TO REPRESENTATIVE DATA SAMPLING

Part 1

Guidelines No. 1

Moscow
2021

THE MOSCOW CENTER
FOR DIAGNOSTICS & TELEMEDICINE

UDC 615.84+616-71
LBC 5c51
M 80

Best Practices in Medical Imaging

Series founded in 2021

Developer:
Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department

Authors:
S.P. Morozov, MD, MPH, PhD, Professor, Chief Regional Radiology and Instrumental Diagnostics Officer, Moscow Health Care Department and the Ministry of Health of the Russian Federation for the Central Federal District; CEO, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
A.V. Vladzymyrskyy, MD, PhD, Chief Research Officer, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Depart
A.E. Andreychenko, PhD, Head of the Department of Medical Informatics, Radiomics, and Radiogenomics, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
E.S. Akhmad, Junior Researcher, Standardization and Quality Assurance Sector, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
I.A. Blokhin, Junior Researcher, Radiology Research Sector, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
V.A. Gombolevsky, MD, PhD, Managing Director of Key Research Programs, Artificial Intelligence Research Institute
V.V. Zinchenko, Head of Clinical and Technical Testing Sector, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
N.S. Kulberg, PhD, Head of the Department of Medical Imaging Tools Development, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
V.P. Novik, Researcher, Department of Medical Imaging Tools Development, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department
N.A. Pavlov, Project Lead, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow Health Care Department

M 80 Dataset preparation regulations and approaches to representative data sampling. Part 1: Guidelines. S.P. Morozov, A.V. Vladzymyrskyy, A.E. Andreychenko et al. Best Practices in Medical Imaging. Issue 103. Moscow. Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, 2021, 40 pp.

Reviewers:
E.I. Kremneva, PhD, Senior Researcher, Department of Radiology, Research Center of Neurology
A.V. Mishchenko, MD, PhD, Deputy Chief Medical Officer, City Clinical Oncological Hospital No. 1, Moscow Health Care Department

This guidance document provides recommendations to all healthcare professionals who curate and perform labeling of medical datasets.

The guidelines were developed as part of the research project "Scientific substantiation for the methodology of application and methods of quality assessment of intelligent technologies (artificial intelligence) in diagnostic imaging".

ISSN 2618-7124

CONTENTS

NORMATIVE REFERENCES

1. GOST ISO 13485-2017 "Quality management systems. Requirements for regulatory purposes".
2. Regulation of the Government of Moscow No. 1543-PP of November 21, 2019 "On conducting the Experiment on the application of innovative computer vision technologies for the analysis of medical images and further use in the Moscow healthcare system".
3. Order of the Moscow Health Care Department No. 51 of January 26, 2021 "On approval of the Procedure and conditions to conduct the Experiment on the application of innovative computer vision technologies for the analysis of medical images and further use in the Moscow healthcare system" (as revised on April 30, 2021 under No. 413, as revised on June 23, 2021 under No. 588).
4. Order of the Federal Service for Supervision of Communications, Information Technology, and Mass Communications (Roskomnadzor) of Moscow No. 996 of September 5, 2013 "On approval of requirements and methods for personal data anonymization".
5. Decree of the President of the Russian Federation No. 490 of October 10, 2019 "On the development of artificial intelligence in the Russian Federation".
6. Federal Law No. 323-FZ of November 21, 2011 "On the fundamentals of health protection of the citizens of the Russian Federation".
7. Federal Law No. 152-FZ of July 27, 2006 "On personal data".

GLOSSARY

The document contains the following terms with appropriate definitions:

1. **Anonymization** (de-identification) refers to removing the connection between identifiable data and the data subject. For this purpose, all attributes are deleted from the record or irreversibly altered in such a way that the data subject can no longer be identified (irreversible de-identification).
2. **Life cycle** refers to the evolution of a system, product, service, project, or other human-made entity from conception through retirement.
3. **AI service** refers to special software based on artificial intelligence (computer vision) algorithms that solves specific medical and diagnostic problems of medical imaging.
4. **Artificial intelligence (AI)** refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones, or Internet of Things applications).
5. **Metadata** refers to information about a dataset, which provides a structured reference that helps to classify, sort, and describe data.
6. **Dataset** refers to an ordered collection of data and corresponding metadata organized according to certain rules.
7. The **reverse engineering process**, i.e. de-anonymization (re-identification), refers to the processing of data in such a way that anonymous data can be attributed to a specific data subject, as a result of which anonymous data becomes personal data.
8. **Pseudonymization** refers to a particular type of de-identification that removes the direct association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.
9. **Data labeling (annotation)** refers to the process of adding data type identifiers to structured and unstructured data (such as text, images, and videos), i.e., data classification and/or data interpretation, in order to solve a specific problem, including by using artificial intelligence technology.

SYMBOLS AND ABBREVIATIONS

The document contains the following symbols and abbreviations:

1. DICOM – Digital Imaging and Communications in Medicine
2. URIS UMIAS – Unified Radiological Information Service of the Unified Medical Information and Analytical System of Moscow
3. AI – Artificial Intelligence
4. FDA – Food and Drug Administration
5. GDPR – General Data Protection Regulation
6. EHR – Electronic Health Record
7. HIS – Health Information System
8. QMS – Quality Management System

INTRODUCTION

The purpose of these guidelines is to describe the main stages of medical dataset creation for machine learning, indicating the required staff resources, tools, and infrastructure.

The guidelines incorporate the practical experience and expertise acquired by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department in the creation of medical datasets for testing, validation, and training of intelligent systems in healthcare, including as part of the Experiment on the application of innovative computer vision technologies for the analysis of medical images and further use in the Moscow healthcare system (Regulation of the Government of Moscow No. 1543-PP of November 21, 2019, Order of the Moscow Health Care Department No. 51 of January 26, 2021 (as revised on April 30, 2021 under No. 413, as revised on June 23, 2021 under No. 588)). The guidelines provide an analytical summary of the world's best practices in the design, creation, and management of medical imaging datasets for artificial intelligence and machine learning. The guidelines reflect all aspects of health data management to ensure that created datasets are fir-for-purpose and meet the requirements of involved stakeholders. These include technical, medical, and regulatory (normative) aspects of preparation and application of datasets for artificial intelligence-based software (hereinafter, the AI services) in healthcare. These guidelines present a general-purpose approach that can be used in any field of healthcare where AI is applicable.

PURPOSE AND RELEVANCE

The diagnosis, treatment, and prevention of diseases based on intelligent analysis of medical big data is one of the promising areas in today's healthcare. For this purpose and to standardize and improve the accuracy of medical interpretation and diagnosis, artificial intelligence-based algorithms and services are being developed. However, the advancement in this field is highly dependent on the availability of high-quality structured datasets collected from medical big data.

The amount of primary digital (machine-readable) electronic health records grows annually and includes both clinical data (medical and physical examination findings) and laboratory and instrumental findings, including diagnostic imaging tests (magnetic resonance imaging, computed tomography, radiography, mammography, fluorography, etc.) and bio-signal analysis (electrocardiography, electroencephalography, electroneuromyography, etc.).

A key limitation for straightforward use of collected data by artificial intelligence (hereinafter, AI) is the fact that electronic health records are compiled during routine medical practice and are not originally indented for data harvesting and machine processing, which leads to unstructured data and different data presentation formats. Although artificial intelligence in healthcare is rapidly developing, the process of health data collection and combining it into a single structure with a set of predefined parameters enabling further manipulations and calculations has not yet been regulated.

Below are some challenges inherent to artificial intelligence in healthcare:

1. Abundance of patient data, including medical images, resulting in the need to regulate the process of data collection to create datasets for machine learning;
2. Emergence of unregistered medical cases that require a prompt response and creation of new datasets to be investigated by clinicians and researchers (the recent example is the detection of the novel coronavirus disease 2019 (COVID-19));
3. Rapid creation and advancement of machine-learning algorithms that require reference datasets for their validation;
4. Reference datasets must have the structure and parameters that enable machine learning applications; moreover, they must be fit-for-purpose both from the computational and medical perspectives. This requires interdisciplinary professionals qualified to collect and prepare requested data.

The purpose of these guidelines is a systematic review of the global practical experience and expertise acquired by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department in the preparation of medical datasets for the development and validation of AI healthcare services. The guidelines will be structured in two parts: Part 1 (this document) comprises methodological issues of medical dataset preparation; Part 2 describes technical details of dataset creation.

Understanding the concepts of datasets and labeling

A dataset refers to a structured set of logically organized and machine-processable data, which has four main features:

1) Content (observations, values, records, files, etc.);
2) Purpose (e.g., knowledge base, particular application);
3) Grouping (aggregation and organization of contents into sets, collections, etc.); and
4) Cohesion (relation to the subject, integration, logical collection of content, etc.).

Data labeling (annotation) refers to the process of adding data type identifiers to structured and unstructured data (such as text, images, and videos), i.e., data classification and/or data interpretation, in order to solve a specific problem, including by using artificial intelligence systems.

Intended use of datasets

Datasets intended for one domain or solving the same clinical/practical task may vary depending on the ultimate goal of their application. The proposed dataset classification is as follows:

1) Self-testing to check for technical readiness of an AI service;
2) Testing on local data for AI service validation and calibration;
3) Transfer learning to retrain an existing AI model;
4) Machine learning to train new models and solve new clinical tasks.

Understanding the concept of dataset lifecycle

The dataset lifecycle model is shown in Figure 1; it provides an overview of the main phases of dataset design, creation, modification, and use. It was adapted from the CRISP-DM model [1] for data mining.

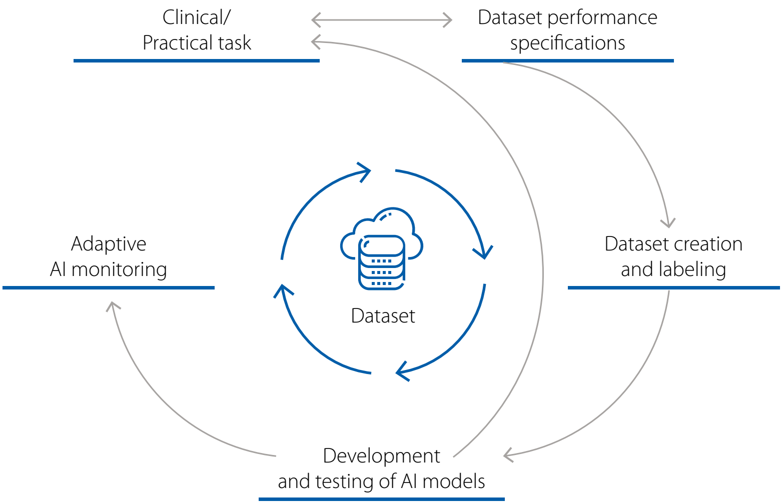

Figure 1 – Dataset lifecycle for artificial intelligence systems in healthcare

While preparing a dataset, the following should be kept in mind: third-party access policy, update frequency, support period, and data destruction (erasure).

Individual types of datasets are subject to regular updates. This may concern both supporting information (e.g., for verification purposes when follow-up studies become available) and dataset units (e.g., when adding new cases in a specific epidemiological context). In these cases, it is possible to additionally describe the principles of new data collection and dataset modification, including changes in the version number.

The dataset can be programmed to change the data access level from closed/restricted to open after a certain while (e.g., one calendar year from the date of publication).

On the contrary, a dataset can have its expiration date, after which it must be either removed from access (from open to restricted/closed access or from restricted to closed access) or archived for long-term storage without the possibility to quickly restore it.

A complete deletion of the dataset is discouraged as, in the future, it may be necessary to restore the source of missing studies.

Understanding the concepts of dataset units and verified datasets

A dataset unit refers to a paired record of inputs and outputs expected from the AI system after processing and analyzing the inputs (Figure 2). The expected outputs are generated during labeling; i.e., as a result of labeling, reference (or "correct") responses are obtained, which will then be used to evaluate the responses from the AI system for its further development, testing and/or post-marketing surveillance.

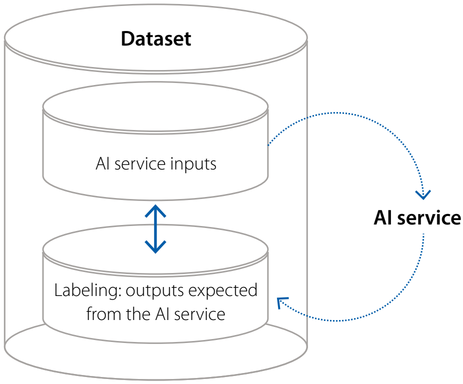

Figure 2 – Dataset sample (the dataset contains paired, unambiguously matching entries of inputs and labeling)

Verification of dataset labeling refers to the audit of dataset labeling results using more accurate diagnostic techniques and methods such as alternative diagnostic tests, clinical diagnosis confirmation (e.g., using biopsy or specific laboratory tests), clinical follow-up results, diagnosis made by the clinician, etc. The verification method depends on the dataset design and the applicable use cases to be addressed by the AI service.

1. HEALTH DATA TYPES AND SOURCES

Health data is broadly defined as any data related to the health status and quality of life of an individual or population. Health data includes clinical metrics along with environmental, socioeconomic, and behavioral information pertinent to health and wellness.

1.1. Health data sources

A lot of health data is collected and used when people interact with health services. Usually, this data is gathered by healthcare providers and includes records related to the services provided, conditions for their provision, and clinical outcomes.

1.2. Classification of types of medical information

The types of medical information are summarized in Table 1 [2].

Table 1 – Common types of health data

Data types	Format description	Main features (challenges of use)
Health records	Data from printed and handwritten texts	Unstructured paper records
Electronic health record	Medical information system for the collection, storage, and display of patient information	Unstructured text
Laboratory data	Software and databases used to manage and store laboratory results and pathology data: in quantitative, qualitative, and graphical form	Lack of standardization in data collection, analysis, and storage, and data access control
Medical images	Medical images are obtained for diagnosis, health status identification, and treatment planning. The most common modalities include PET, CT, CBCT, MRI, and ultrasonography. Medical imaging is regulated by the generally accepted DICOM standard	Inadequate compliance with standardization guidelines for data collection and analysis; data duplication within a single medical facility; data availability
Genomic data	Individual datasets with large-scale genomic data	Incomplete data; data availability
Auxiliary data	Income, social status, race, ethnicity, education, housing	Unstructured and incomplete data; data availability

Laboratory data: Most of the data can be in the form of numerical or categorical values. In addition, there is a separate group of pathological tests, which include:

1) Electron microscopy;
2) Whole mount staining;
3) Permanent mount technique;
4) Temporary mount technique;
5) Tissue culture technique;
6) Autoradiography.

Such examinations can also be presented as images in various formats. When preparing a dataset, it is necessary to choose one image format and one format of supporting documentation.

Medical images are the images of internal structures of the human body intended for clinical analysis and medical interventions as well as for visual representation of the functions of certain organs or tissues, which are obtained noninvasively by special devices and sensors. Medical images of the following common modalities are stored in the DICOM format:

EPS – Cardiac Electrophysiology;
CR – Computed Radiography;
CT – Computed Tomography;
DX – Digital Radiography;
ECG – Electrocardiography;
ES – Endoscopy;
XC – External-camera Photography;
IVUS – Intravascular Ultrasound;
MR – Magnetic Resonance;
MG – Mammography;
NM – Nuclear Medicine;
OP – Ophthalmic Photography;
PX – Panoramic X-Ray;
PT – Positron emission tomography;
RF – Radiofluoroscopy;
RG – Radiographic imaging;
US – Ultrasound;
XA – X-Ray Angiography;
BI – Biomagnetic Imaging;
CD – Color flow Doppler;
ST – Single-Photon Emission Computed Tomography (SPECT);
TG – Thermography;
AU – Audio;
SR – SR Document;
SMR – Stereometric Relationship;

SC – Secondary Capture;

OT – Other.

Normally, medical images are stored in the DICOM format; however, sometimes it cannot store the full amount of diagnostic data (e.g., for spectral images).

After image processing, the data can be converted to other storage formats such a NIFTI, various graphic formats (jpg, png), and others.

2. APPROACHES TO THE SELECTION OF A USE CASE TO BE ADDRESSED BY ARTIFICIAL INTELLIGENCE IN HEALTHCARE

An ideal AI use case should be specific, measurable, and achievable, and have well-defined users and value [3]. When defining the use case, one can be guided by the adapted SMART-GEM scale [4]. Use cases also help illustrate reporting guidelines or the need to develop them [5].

AI algorithms in healthcare create new value for clinicians and patients as they help healthcare providers improve patient treatment outcomes and reduce healthcare costs [6]. However, commercially available AI algorithms for medical image analysis require independent evaluation and potential approval by the Food and Drug Administration (FDA) [7]. Datasets for independent validation of AI algorithms should include data elements that provide adequate understanding of limitations, biases, and even potential ethical concerns caused by the use of a specific AI service [8].

An AI use case in healthcare environments should reflect its clinical task, including clinically important outcomes based on the current clinical guidelines and findings easily amenable to independent human identification [9]. However, AI can be misused, for example, in case of applying a diagnostic screening algorithm [10] (e.g., intended for tuberculosis screening) in urgent care. This is incorrect as the algorithm is designed to detect infection signs in apparently healthy individuals, and its use in emergency settings to detect hemothorax, pneumothorax, hydrothorax, or other urgent conditions may pose a risk for patients' life as such abnormalities are out the scope of screening tests.

Thus, an AI algorithm for use in healthcare environments should demonstrate [11]:

1) Clinical utility, to improve clinical care;

2) Statistical validity, by training the model on large-scale and diverse datasets to achieve high reliability in a new population; and

3) Economic utility, as shown in prospective or retrospective settings.

3. APPROACHES TO DATASET CREATION

3.1. Preparation of performance specifications

Figure 3 shows the main stages of dataset creation to be addressed in these guidelines.

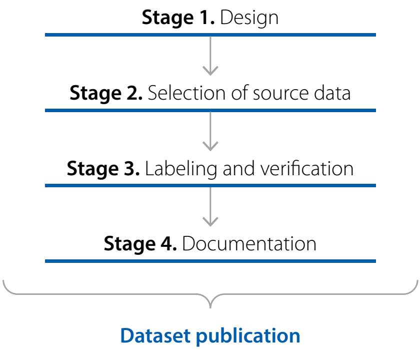

Figure 3 – Stages of dataset creation

An important stage of dataset creation is the choice of its design, i.e., the articulation of functional requirements for dataset performance (performance specifications). This stage is necessary to specify the technical requirements for a dataset in strict correspondence with its clinical and practical tasks.

3.1.1. Setting clinical and practical tasks of dataset creation

A clinical or practical task means the AI use case in medical imaging, differential diagnosis (independent reading), decision support (increasing the range of inputs for decision-making by a healthcare professional), patient routing (triage), technical support (reducing the burden of routine tasks), etc.

The task meets the relevance criteria if:

1) It covers the needs of medical community; in other words, healthcare professionals understand the outcomes they will get from the AI services and are interested in them;

2) Corresponding AI solutions are marketed both in Russia and abroad; and

3) Solution of this task will bring about a significant socioeconomic effect.

To be considered relevant, it is not mandatory for a task to meet all the free criteria; for instance, an automation task may be relevant even if no AI solutions are available on the market but the task is strategically important for healthcare.

When setting clinical and practical tasks, it is necessary to define the data access and processing rules and specify:

1) What kind of data may be collected, in what amount, and from which source;

2) How it should be used (with respect to specific tasks);

3) How data will be protected during AI service operation;

4) To whom it should be disclosed (third-party access policy); and

5) For how long it should be available.

3.1.2. Defining the dataset parameters

It is necessary to define the expected AI service performance results and dataset requirements (including labeling and verification requirements). This process involves researchers, clinicians, and biostatisticians.

Paragraphs 3.1.2.1 – 3.1.2.7 of this section provide the basis for completing the dataset performance specifications. As a result, dataset performance specifications and baseline requirements for AI service performance will be defined.

3.1.2.1. Characteristics of clinical and practical tasks

The clinical task may be dichotomous (division into two classes, such as "signs of target disease entity present/absent" – in this case, the goal of binary classification is achieved) or it may contain a differential diagnosis (such as the differentiation of one pathology from another, i.e., multiclass classification). The clinical task may also consist in measuring a continuous variable (e.g., the percentage of parenchymal involvement) or in the detection, search, and display of imaging findings, as well as in choosing patient management strategy, analysis-based prediction of outcomes, etc.

The requirements for the AI outputs may vary depending on the clinical task, namely:

1) A choice between one of the classes or a probability of one of the classes in percent – for binary classification;

2) A choice of the most probable class or a probability of each class in percent – for multiclass classification;

3) Prediction of the variable value indicating the known error – for a continuous variable;

4) Coordinates of the finding, heatmap, or contours – for imaging findings.

3.1.2.2. Defining the dataset scope of application

Such parameters as class balance (equal number of dataset units for different classes, for example, norms and pathologies for binary classification, see para. 3.1.2.6), target number of studies, etc. may vary depending on the dataset scope of application.

At this stage, the dataset intended use should be kept in mind, in particular:

1) AI service testing for its validation and verification:

– Functional testing (control of the operability of algorithms, visual assessment of outputs, etc.);

– Clinical validation (verification of the AI service accuracy metrics considering AI service use cases);

– Self-testing (independent verification of the AI service ability to process heterogeneous input data carried out by the developer);

2) Machine learning:

– Transfer learning;

– Algorithm training.

In the case of using the dataset for machine learning, it is important to consider the AI application pattern. For instance, to develop an AI service that will be used "before the clinician's review", i.e., screen for "normal" studies without abnormalities, the dataset must contain anatomic variations. In contrast, the dataset for the development of an AI service that will be used "by the clinician" should be aimed at differential diagnosis, complex cases, etc.

3.1.2.3. Defining clinical parameters of the dataset

To describe each dataset, it is necessary to provide a check list that specifies all clinical parameters of the dataset in relation to a specific healthcare domain.

The proposed version of the check list of clinical parameters is applicable for diagnostic investigations:

1) Type of input data according to generally accepted standards (e.g., for medical images – according to the DICOM standard [12]);

2) Anatomical location (according to the normative reference data [13]);

3) Target disease entity (one or more, according to the normative reference data [14]);

4) Population criteria:

- Specified upper age limit;
- Specified lower age limit;
- Gender distribution;
- Location and dates of data collection;

- Characteristics of a medical facility taking part in data collection:
– Name;
– Type;
– Type of care (pediatric, adult, mixed);
- Epidemiological situation during data collection;
- Other patient selection criteria.

3.1.2.4. Defining technical parameters of the dataset

It is necessary to specify the following technical parameters of the dataset:
1) Characteristics of diagnostic devices the data was obtained from:
– List of manufacturers (and models, if necessary);
– Technical specifications;
– Availability of special imaging modes;
2) Data requirements (e.g., resolution);
3) Requirements for de-identification:
- De-identification of metadata:
– According to generally accepted standards (e.g., the DICOM standard, Section E1, Table E.1-1 [15] – for medical images);
– Preserving specific personal information for future comparison with the supporting materials;
- De-identification of pixel images:
– Detection of text in images;
– Removal of facial soft tissues for head, neck, and brain imaging.

3.1.2.5. Defining the labeling criteria

Labeling criteria are a prerequisite for proper labeling. They include the following:
1) Dataset inputs:
– Name of the input data unit;
– Format of the input data unit;
2) Dataset outputs:
– Name of the output data unit;
– Format of the output data unit;
3) Labeling classification:
– Single-label;
– Multi-label;
4) For each label:
– Level of labeling (from the list):

- Patient;
- Study;
- Series;
- Image;
- Level of labeling details:
- Study/series/image;
- Finding (location);
- Finding (segmentation);
- Type of label:
- Binary classification;
- Multiclass (more than 2 dependent classes);
- Continuous variable;
- For each class:
- Class inclusion criteria;
- Class exclusion criteria;
- Data source for criteria (the study itself (images), metadata, other sources);
- Literature reference.

3.1.2.6. Defining the class balance and the target number of studies

To define the class balance and the target number of studies, it is necessary to consider the scope of application of a dataset and the degree of labeling complexity (number of labels and classes).

Classes of one label can be:

1) Balanced (the number of studies is the same across different classes);

2) Imbalanced (the number of studies of one class is prevalent).

One of the most common imbalanced datasets are the datasets based on the pre-test probability. In this case, the number of studies with abnormal signs will be comparable to the detected number of such signs in a given population.

To achieve the goals of testing by receiver-operating characteristic analysis (ROC analysis), balanced datasets are needed.

The target number of studies of each class can be calculated statistically based on the required power of the test. To calculate the dataset size, biostatistical methods should be applied.

3.1.2.7. Defining the sources of dataset source data

Typically, source data is harvested from medical information systems (e.g., URIS UMIAS in Moscow); however, data can come from other databases and physical carriers as well; supporting documentation can also be downloaded from other sources (such as clinical diagnosis data, etc.).

3.2. Source data collection pursuant to performance specifications

3.2.1. Introduction to digital health data

Rationale for the AI development, data preparation for labeling and testing, and prospective AI performance depend strongly on the access to source data, which can be either raw or pre-processed using medical software and is available to the end user (healthcare professionals). Raw data mean unprocessed data collected from diagnostic devices, which is often inaccessible both to the user and medical information systems, and represent a complex mathematical dataset that has no value for the clinician. Raw data is valuable for software developers who fine-tune the algorithms for medical signal preprocessing.

3.2.2. Regulatory framework for the collection of source data

Access to source data is restricted by the following regulations:

1. Federal Law No. 323-FZ of November 21, 2011 "On the fundamentals of health protection of the citizens of the Russian Federation" (as amended effective from July 13, 2021): Article 4; Article 13, Parts 2–4; Article 92.
2. Federal Law No. 152-FZ of July 27, 2006 "On personal data". Excerpts related to the collection, processing, and transmission of datasets: Articles 5 and 6.

Under the European General Data Protection Regulation (GDPR) [16], personal data includes all data which are or can be assigned to a person directly or indirectly, which is a broader interpretation compared to the one accepted in the Russian Federation (pursuant to Federal Law No. 152-FZ of July 27, 2006 "On personal data").

De-identification refers to the actions that result in inability to identify a specific data subject without the use of additional information. The main purpose of de-identification is to ensure that personal data is kept confidential.

Normative and methodological regulations that address the issue of removing the connection between personal data and data subject use three common terms to define this process:

1. Anonymization (de-identification) refers to removing the connection between the identifiable data and the data subject. For this purpose, all attributes are deleted from the record or irreversibly altered in such a way that the data subject can no longer be identified (irreversible de-identification).
2. Pseudonymization refers to a particular type of de-identification that removes the direct association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms. This process is reversible. As for anonymization, attributes are altered or deleted from the record, but anonymized patient data is associated with a pseudonym.

3. The reverse engineering process, i.e. de-anonymization (re-identification), refers to the processing of data in such a way that anonymous data can be attributed to a specific data subject, as a result of which anonymous data becomes personal data.

When creating datasets, the terms "anonymization", and "de-identification" are synonymous.

Pursuant to Order of the Federal Service for Supervision of Communications, Information Technology, and Mass Communications (Roskomnadzor) of Moscow No. 996 of September 5, 2013 "On approval of requirements and methods for personal data anonymization", anonymized data is:

1) Exhaustive (it has all information about specific subjects or groups of subjects that was available before anonymization);

2) Structured (it retains structural links between the de-identified data of a specific subject or a group of subjects matching those links that existed before anonymization);

3) Relevant (is enables the processing of personal data in such a way that the request and responses have the same semantic form);

4) Characterized by semantic integrity (after anonymization, is preserves the semantics of personal data);

5) Applicable (it provides the possibility of achieving personal data processing goals by the operator who de-identifies personal data handled in personal data information systems, including data created and managed in federal target programs (hereinafter, the operator(s)), without prior de-anonymization of the entire volume of subject records);

6) Anonymous (after de-identification, it is impossible to unambiguously identify data subjects without the use of additional information).

Thus, the collection of source data is highly restricted by the current regulations that must be observed when translating AI into clinical care.

3.2.3. Preparing the infrastructure for source data harvesting

At the initial stage, software solutions are developed to automate the preparation of source data for labeling, to provide labeling itself, and to create datasets. The work at this stage involves a software development team, including a systems architect, database architect, user interface designer, DevOps engineer, programmers, and testers. To obtain source data from medical information systems, single-unit or batch data processing may be applied. To prepare datasets intended for AI applications, batch data processing is preferable as, typically, large amounts of studies from different patients are needed. It requires software with appropriate functionality. As shown by the example of digital imaging data from medical facilities of the Moscow Health Care Department, the software should incorporate the following functionality:

1) Search for source data in URIS UMIAS;

2) Download text reports using available lists of unique identifiers of studies (study_uids);

3) Select studies by keywords (pre-sorting);

4) Select and filter studies by technical parameters;

5) Ensure first/second reading by radiologists (annotators and experts);

6) Ensure verification;

7) Save labeling and verification results in machine-readable form;

8) Prepare dataset supporting documentation.

3.2.4. Source data collection

An important feature when planning the source data collection is its availability, which depends on the data source (Table 2).

Table 2 – Data sources

Data source types	Data types
Medical sources	– Images (CT, MRI, surgery videos, etc.) – Text (EHR, medical reports and guidelines, etc.) – Sounds (patient voice recordings, wheezing and coughing sounds, etc.) – Signals (EEG, ECG, data from bedside monitors and wearable devices, etc.) – Genetic data (NGS, DNA microarray, etc.)
Pharmacological sources	– Clinical trials data – Data on medicines – Sales data – Pharmacovigilance data
Financial sources	– Cash flow data from the Compulsory Medical Insurance Fund and medical facilities – Cash flow data from insurance companies
Administrative sources	– Data on medical facilities (doctor's office workload, equipment load, working hours, etc.) – Federal Register of Medical Facilities (FRMO), Federal Register of Healthcare Professionals (FRMR), Normative Reference Data (NSI) – Data on complaints and feedback on medical facilities – Data on disease prevalence, epidemics, etc.
Insurance data	– Insurance reports – Customer scoring reports – Complaints data

External sources	– Demographic data – Biomedical literature – Data from patient forums – Cancer database (cancer registry) – Database of the Registry Office – Open-access databases

When additional characteristics are added to the designed dataset, the number of subjects who will simultaneously have the entire set of such characteristics decreases; therefore, the planning stage of source data collection should give the understanding of each dataset purpose.

Since 2013, the Ministry of Health of the Russian Federation has regulated the structure of an electronic health record (hereinafter, the EHR) [17] to be applied when creating and improving medical information systems (hereinafter, the MIS).

The EHR enables a long-term storage of patient data related to all types of medical care, including the results of medical follow-up, clinical judgments, and treatment plans.

The structure of the EHR includes 15 sections such as "Patient metrics", "Diagnostic tests", "Medical examinations", "Diseases and complications", "Medications", and others.

Sections, in turn, include dozens of parameters (EHR fields); for instance, the "Medical examinations" section must contain the full name and position of doctor, symptoms and complaints, diagnosis, etc. Structured reporting facilitates more accurate query to collect various datasets.

3.2.5. Stages of dataset creation

Most healthcare systems are not adequately equipped to share large amounts of medical images [18]. Even when ethical, regulatory, and financial prerequisites for AI development/testing/application are fulfilled, rapid proliferation of AI in clinical care is hindered as health data is often stored in disparate repositories. Collection of large amounts of source data from one place prevents it from being used for appropriate development, testing and/or post-marketing surveillance of medical AI. Furthermore, the process of collecting and storing source data lacks sound methodology, which should be elaborated before handling the source data. One of the possible solutions is to transfer the responsibility for and rights to the creation, training, and testing in the common digital environment from one medical facility to another. This is the case in Moscow, where the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department took over the responsibility for the methodology and implementation of AI in the Experiment on the application of innovative computer vision technologies for the analysis of medical images and further use in the Moscow healthcare system [19].

To prepare source data, the following steps need to be taken:

1. Planning the goals of dataset application. This important step is often neglected, but it defines the number of patients/studies, dataset scope of application; inclusion/exclusion criteria; availability of ground-truth labels in the dataset; the need (or possibility) to update (or expand) the dataset to extend the scope of application of available data, including limitations of dataset use.
2. Approval from the ethical committee to use the source data for a specific purpose. To use source data, an informed consent is required.
3. Getting access to the desired source data using relevant query.
4. Data anonymization and secure storage of de-anonymization (re-identification) keys.
5. Data quality control. The quality and amount of images vary depending on the target task and domain. If the data is intended for open-source research, then additional human inspection of each image is standard because some images contain free-form annotations that cannot be removed reliably with automated methods.
6. Data structuring in homogenized and machine-readable formats [20] (e.g., DICOM or NIFTI).
7. Linking the images to ground-truth information, which can be one or multiple labels, segmentations, or electronic phenotype (e.g., biopsy or laboratory results).
8. Registration of the dataset as an independent intellectual property.

3.2.6. Recommendations for data collection

Depending on the domains the data may be used in for training, transfer learning, testing, validation, or scientific analysis, datasets can be collected in the following ways:

1) Retrospective/Prospective data collection (must follow the above steps);

2) Research efforts;

3) Using publicly available databases with appropriate authorization, which is the easiest way that has its limitations, namely:

- Some open-source datasets may be used for research and not for the commercial AI development;
- It is impossible to ensure the dataset quality control;
- The dataset size is limited.

The search for ground-truth information to confirm the target pathology in medical images is a challenging and topical problem. In addition to the image labeling, which can be very time-intensive, each study should be interpreted in line with the corresponding reporting guidelines. Their use can either result in image labeling or contribute to reducing the number of images that will require labeling in the future. There are approaches to perform retrospective labeling, ranging from simple manual labeling [21] by radiologists to automated approaches that can extract structured information from the radiology report and/or electronic health record.

There is a trend toward interactive reporting where the radiologist report contains hypertext directly connected to image annotations [22]. Such annotations can be used for labeling of open-source datasets [23]. However, this approach cannot be considered reliable as 2–20% of radiology reports contain errors [24].

When it comes to image annotation, priority should be given to medical professionals with sufficient experience, keeping in mind that an expert radiologist should validate such labeling as an auditor. Such crowd-sourced labeling should be performed after a preliminary discussion with experts since the number of annotators and the need for an auditor may vary depending on the task. Thus, complex tasks require a substantial number of annotators; for instance, the recommended number of experts to label lung lesions is 4 labelers and 1 expert validator for each CT scan [25].

When planning data collection, it is advisable to design the dataset in such a way that the ratio of training, testing, and validation in the dataset is 80:10:10 or 70:15:15. To ensure generalizability of the AI algorithm, bias of the training dataset should be limited. If the AI algorithm is trained on images from a Moscow facility and the algorithm is used in the Asian population, its performance may be affected by population bias or prevalence bias. Similarly, if all the imaging training data was acquired by using one kind of imaging machine, it may not work as well on machines from other manufacturers. It is thus advised to use images from multiple diverse sources, or at least images representing the target population or health system in which the algorithm is to be deployed. To ensure generalizability, large training datasets are often essential. For specific targeted applications or populations, relatively small datasets (hundreds of cases) may be sufficient. Large sample sizes are especially required in populations with substantial heterogeneity or when differences between imaging phenotypes are subtle [26].

The sample size calculation for test datasets should use traditional power calculation methods to estimate the sample size. In general, the development of generalizable AI algorithms in medical imaging requires statistically powered datasets in the order of hundreds of thousands, which is problematic for many researchers and developers. A partial solution for this problem may be semi-supervised learning. Fully annotated datasets are needed for supervised learning, whereas semi-supervised learning [27] uses a combination of annotated and unannotated images to train the algorithm.

Special attention should be given to federated learning. A number of companies [28, 29] enabled multi-institutional collaborations in order to train AI models using in-house computation resources without sharing data between institutions. As the trained models are exchanged between medical institutions, they are fine-tuned and boast higher diagnostic accuracy. Training, testing, and prospective work are carried out inside the institution without the need to transfer datasets (studies) outside the federation. Despite the potential benefits of this technique, there are important problems that need to be solved before federated learning can be widely applied in practice:

1. It is necessary to standardize the interpretation and labeling of medical data in those institutions where the medical AI algorithm is to be deployed.
2. Depending on the AI algorithm complexity, substantial computational resources need to be placed within each institution for federated learning.
3. Preprocessing and organizing the data for ingestion by the AI algorithm is challenging, because the visibility of data to the developers is impeded.
4. Data heterogeneity across different institutions in terms of patient populations, pathology distribution, data volume, data format, etc.

One of the most important limitations of training AI algorithms based on data from a single institution or from multiple institutions in a small geographic area is sampling bias. If an AI algorithm trained this way is applied to a different geographic area, then results of the algorithm may be unreliable due to differences between the sample population and target population [30].

Consequently, there are various recommendations for data collection and labeling for AI. Clinicians, data scientists, and decision-makers who provide access to data or authorize new AI model deployment should be aware of the source for training data and potential biases, which may affect generalizability of AI algorithms. New approaches such as federated learning, interactive reporting, and synoptic reporting may help to address the problem of data availability in the future. However, curating and annotating data, as well as computational requirements, are substantial barriers.

3.3. Dataset classification by the type of labeling

The type of image labeling varies depending on the task to be performed by the AI algorithm [31].

There are three types of datasets depending on the type of labeling (Fig. 4, 5) [18]:

1) Retrospective dataset;
2) Prospective dataset;
3) Verified dataset.

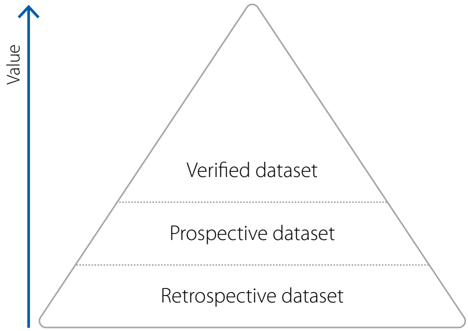

Figure 4 – Value-based classification of image labeling

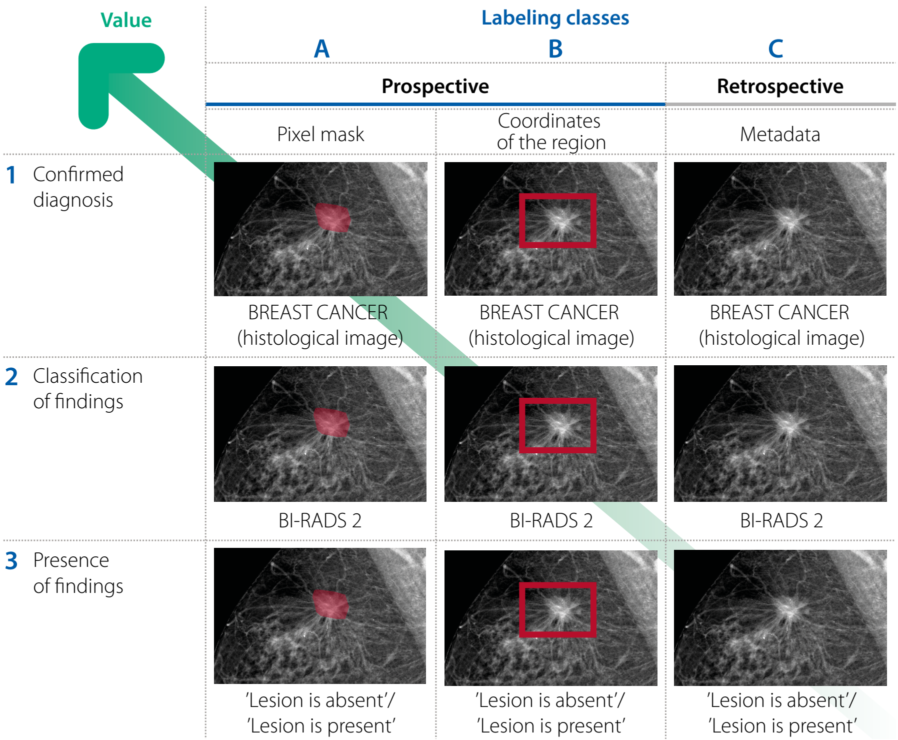

Figure 5 – Labelling classification in medical imaging

Figure 5 above shows the classification of labeling types [32]. For retrospective labeling (1C, 2C, 3C), data from supporting documentation (e.g., radiology reports), medical information systems, electronic health records, etc. can be used. An example of retrospective labeling is metadata generated automatically during the examination, which is stored in the source data. An obvious advantage of retrospective labeling is its time-efficiency as it is the data scientist, not the doctor, who does most of the preparations.

A dataset is considered most reliable (line 1) when a confirmed diagnosis is available for elements in the dataset, e.g., the results of histological tests, laboratory data, or follow-up studies (if applicable).

Prospective labeling (1A, 1B) requires the doctors to be actively involved in the process of dataset creation to include information for efficient classification of dataset elements. In radiology, labeling is most often understood as the classification of studies by classes (i.e., depending on the presence or absence of radiological signs of a specific disease), as well as segmentation of the region of interest [18] that has the required signs of pathology (e.g., demyelinating lesions in multiple sclerosis in brain MRI images). The doctor involvement can be more or less time-intensive: in the first case, experts are to outline the region of interest with a contour, i.e., create a pixel mask at the level of the region of interest (Column A), in the second case, they are to designate its coordinates with a simple geometric figure (Column B).

3.3.1. Retrospective labeling

Retrospective labeling enables the collection of elements corresponding to metadata selected in line with the set task. Such labeling is not labor-intensive: it consists in uploading data from medical information system, which can be done by an engineer (analyst) without doctor involvement. For this purpose, each element (image, signal data, etc.) of the dataset is linked to health data (diagnosis, laboratory results, etc.).

3.3.2. Prospective labeling

Similar to retrospective labeling, prospective labeling enables the collection of elements in accordance with the set task, but here additional manipulations with elements are mandatory (e.g., providing annotations for event start/end, detection of signs and abnormalities, etc.). This type of labeling involves healthcare professionals (normally, a doctor qualified in the domain of the labeled dataset) who perform manual annotation of data in full or in part.

3.3.3. Verified dataset

The verified dataset is collected by supplementing the dataset prepared during prospective labeling by doctors with data from medical records, including the final (clinical) and/or pathologic diagnosis. The golden standard of dataset verification for the target pathology is getting the ground truth from follow-up examinations, histopathologic, immunologic, and other tests, treatment responses, etc. [18, 20, 31, 33].

There is one more verification technique where the dataset is evaluated by blinded medical experts with the given level of consistency. When multiple expert readers are involved in the dataset verification, it is necessary to describe the process by which their interpretations are combined to make an overall reference standard determination and how the process accounts for any inconsistencies between clinicians participating in the truthing process (ground truth variability) [34]. The difference from prospective labeling is that data is reviewed by a panel of experts who provide a consensus decision.

The dataset is considered verified if:

1) It contains real-practice data (gathering of synthesized data, e.g., from an ECG waveform generator, is prohibited);

2) Dataset structure is consistent with the dataset purpose (training, analytical validation, clinical validation, etc.);

3) The number of observations (studies) ensures the statistical significance of the result;

4) Labeling is done by the group of experts;

5) A thesaurus (i.e., a coded library of words and phrases adopted by the clinical societies' guidelines) is used for labeling.

3.3.4. Requirements for dataset annotators

As per GOST ISO 13485, personnel (doctors, engineers) preparing the dataset must be competent on the basis of appropriate education, training, skills, and experience. Details on the required qualification, experience, and skills of personnel must be specified in their job descriptions.

Annotators should meet the following selection criteria:

1. They should be competent with respect to specific data types: images, text or signal data (ECG, EEG, etc.), quantitative data (heart rate, blood pressure, spirometry parameters, etc.), and binary data (e.g., Yes/No response).

2. Depending on the complexity of the required labeling and/or annotation, it may include primary labeling (segmentation) or expert labeling; providing details at the level of classes or subclasses, establishing links with metadata, or predicting possible outcomes (forecasting).

At that point, the signs of common disease entities (such as pneumonia or tubular bone fractures) do not deserve the attention of a highly-qualified radiologist who, however, should be engaged in the annotation of complex cases and differential diagnosis (e.g., annotation of images with the signs of demyelinating disorders).

3.4. Dataset quality control

Dataset creation process is subject to planning, monitoring, and management to ensure quality compliance.

The working group may be led by a responsible employee who is not engaged in the labeling/annotation process, but will manage the project goals depending on their urgency, priority, and workload of experts. It is responsibility of this employee to form a working group to ensure the fairness and reliability of outcomes.

Dataset quality assurance should be applied to ensure that:

1) The dataset has no missing elements;

2) All the dataset elements are consistent with the set tasks;

3) The quality of the dataset elements complies with the community guidelines.

To develop and apply the verified dataset, a quality management system (QMS) is implemented, which is expressed as the organizational goals and aspirations, policies, processes, documented information, and resources needed to implement and maintain it.

The prepared datasets can be structured by segmenting features in accordance with the set task. In the process of structuring, the dimensionality of the dataset is reduced, leaving a sufficient list of attributes for an accurate and complete description of the dataset elements, which will facilitate the subsequent generalization of steps and high-quality labeling (annotation) of data.

Dataset filtering reduces the annotation costs by excluding data that falls short of the specified parameters. The quality control procedure includes finding, preventing, and eliminating problems related to the quality of datasets. Filtering and quality control of datasets can be done by visual control, special tools (e.g., DICOM validators), or an artificial intelligence system (e.g., for automatic image quality assessment).

3.5. Modification of datasets

After creating and registering a dataset, it may be necessary to modify it (to correct errors or add new data) [35]. When making any changes to the dataset, any changes (including change in the version number) should be documented in order to be able to assess them later. Such documentation must be enclosed to the dataset.

When changing the dataset version, three-digit values are used in the A.B.C format, where A is the major version, B is the minor version, and C is the patch version [32]:

1. The major version increases if there are changes in the meaningful parameters of the dataset related to the clinical task, purpose, principles of data labeling and verification.
2. The minor version increases if data units (images, text, or signal data, etc.) are replaced, added, or deleted without changing the meaningful parameters of the dataset (the minor version is set to 0 when a new major version is released).
3. The patch version increases if changes are made to supporting documentation or typos and errors are corrected in the labeling and verification files, but neither the amount nor the quality of dataset elements providing inputs to the AI service change (the patch version is set to 0 when a new minor or major version is released).

When a new minor or patch version is released, AI services can use the dataset without changing the code that ingests dataset elements that provide inputs. When a new patch version of the dataset is released, the amount and quality of dataset elements providing inputs to the AI service should be the same, but the performance results may be different (because labeling and verification files may be affected). When an additional series of data units is added, the major version increases since meaningful changes are introduced to the clinical task and overall purpose of the dataset creation, which, however, do not change it completely. A new dataset is created if the intended use, purpose, and clinical tasks of the dataset creation are changed completely.

CONCLUSION

These guidelines describe practical approaches to the design and creation of medical datasets intended for the testing and development of artificial intelligence technology in healthcare. Incorporation of these recommendations into routine practice will standardize the development of medical datasets to ensure their value for artificial intelligence systems in healthcare. As the development and validation of useful, reliable, and secure artificial intelligence systems is preconditioned by the availability of high-quality datasets, it is the process of their creation that requires a transparent and reproducible methodology.

REFERENCES

1. Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing. 2000. No. 5, pp. 13-22.

2. Kazmierska J., Hope A., Spezi E. et al. From multisource data to clinical decision aids in radiation oncology: The need for a clinical data science community. Radiotherapy and Oncology. 2020. No. 153, pp. 43–54.

3. Kohli M.D., Summers R. M., Geis J. R. Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session. Journal of Digital Imaging. 2017. Vol. 30, No. 4, pp. 392-399.

4. Bowman J., Mogensen L., Marsland E. et al. The development, content validity and inter-rater reliability of the SMART-Goal Evaluation Method: A standardised method for evaluating clinical goals. Australian occupational therapy journal. 2015. No. 62.6, pp. 420-427.

5. Sounderajah V., Ashrafian H., Aggarwal R. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nature medicine. 2020. Vol. 26, No. 6, pp. 807-808.

6. Lee D.H., Yoon S.N. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. International Journal of Environmental Research and Public Health. 2021. Vol. 18, No. 1, p. 271.

7. Wu E., Wu K., Daneshjou R. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nature medicine. 2021. Vol. 27, No. 4, pp. 582-584.

8. O'Reilly-Shah V., Gentry K.R., Walters A.M. et al. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. British Journal of Anaesthesia. 2020. Vol. 125, No. 6, pp. 843-846.

9. Oren O. B., Gersh J., Bhatt D. L. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. The Lancet Digital Health. 2020. No. 2.9, pp. e486-e488.

10. Melendez J., Sánchez C.I., Philipsen R.H.H.M. et al. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Scientific Reports. 2016. No. 6:25265, pp. 1-8.

11. Sendak M.P., D'Arcy J., Kashyap S. et al. A path for translation of machine learning products into healthcare delivery. EMJ Innov. 2020. No. 10, pp. 1-14. DOI: 10.33590/emjinnov/19-00172.

12. DICOM Library - Anonymize, Share, View DICOM files ONLINE: [website]. Poland, 2021. URL: https://www.dicomlibrary.com/dicom/modality/ (accessed on: July 3, 2021).

13. Normative reference data: [website]. Russia, 2021. URL: https://nsi.rosminzdrav.ru/#!/refbook/1.2.643.5.1.13.13.11.1477 (accessed on: July 3, 2021).

14. Normative reference data: [website]. Russia, 2021. URL: https://nsi.rosminzdrav.ru/#!/ (accessed on: July 3, 2021).

15. DICOM: [website]. United States, 2021. URL: http://dicom.nema.org/dicom/2013/output/chtml/part15/chapter_E.html (accessed on: July 3, 2021).

16. General Data Protection Regulation (GDPR) – Official Legal Text: [website]. Germany, 2021. URL: https://gdpr-info.eu/General Data Protection Regulation (accessed on: September 3, 2021).

17. Ministry of Health of the Russian Federation. Main sections of the electronic health record: [website]. Russia, 2021. URL: https://nsi.rosminzdrav.ru/#!/ (accessed on: July 3, 2021).

18. Willemink M.J., Koszek W.A., Hardell C. et al. Preparing Medical Imaging Data for Machine Learning. DOI: 10.1148/radiol.2020192224. Radiology. 2020. Vol. 295, No. 1, pp. 4-15.

19. Pavlov N., Kirpichev Y.S., Revazyan A. et al. Value of technical stratification of medical datasets for AI services. DOI: 10.1186/s13244-021-01014-5. Insights Imaging. 2021. No. 12 (Suppl 2): 75, p. 216.

20. Harvey H., Glocker B. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. DOI: 10.1007/978-3-319-94878-2_6. Artificial Intelligence in Medical Imaging. 2019, pp. 61-72.

21. Wang X., Peng Y., Lu L. et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 2097-2106.

22. Folio L.R., Machado L.B., Dwyer A.J. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. RadioGraphics. 2018. Vol. 38, No. 2, pp. 46-482.

23. Yan K., Wang X., Lu L. et al. Deep Lesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging. 2018. Vol. 5, No. 3, pp. 1–11.

24. Brady A., Laoide R.O., McCarthy P. et al. Discrepancy and error in radiology: concepts, causes and consequences. The Ulster medical journal. 2012. Vol. 81, No. 1, pp. 3–9.

25. Kulberg N.S., Reshetnikov R.V., Novik V.P. et al. Inter-observer variability between readers of CT images: all for one and one for all. Digital Diagnostics. 2021. Vol. 2, No. 2, pp. 105–118.

26. Chang K., Balachandar N., Lam C. et al. Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association. 2018. Vol. 25, No. 8, pp. 945–954.

27. Kingma D.P., Rezende D.J., Mohamed S. et al. Semi-Supervised Learning with Deep Generative Models. 2014, pp. 1-9. URL: https://arxiv.org/abs/1406.5298 (accessed on: August 11, 2021).

28. Sheller M.J., Edwards B., Reina G.A. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports. 2020. Vol. 10, No. 1, p. 12. DOI 10.1038/s41598-020-69250-1.

29. McMahan B., Ramage D. Federated Learning: Collaborative Machine Learning without Centralized Training Data. Google AI Blog: [website]. 2017. Apr. 6. URL: https://www.https://ai.googleblog.com/2017/04/federated-learning-collaborative.html (accessed on: August 9, 2021).

30. Toll D.B., Janssen K.J., Vergouwe Y. et al. Validation, updating and impact of clinical prediction rules: a review. Journal of clinical epidemiology. 2008. Vol. 61, No. 11, pp. 1085-1094. URL: DOI 10.1016/j.jclinepi.2008.04.008.

31. Diaz O., Kushibar K., Osuala R. et al. Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. Physica Medica. 2021. Vol. 83, No. 5, pp. 25-37.

32. Pavlov N.A., Andreychenko A.E., Vladzymyrskyy A.V. et al. Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics. Digital Diagnostics. 2021. Vol. 2, No. 1, pp. 49-66.

33. Ranschaert E.R., Morozov S.P., Paul R. Artificial Intelligence in Medical Imaging. Opportunities, Applications and Risks. Artificial Intelligence in Medical Imaging. 2019, p. 705.

34. U.S. Food and Drug Administration. Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in - Premarket Notification (510(k)). Submissions Guidance for Industry and FDA Staff: [website]. Netherlands, 2021. URL: https:www.fda.gov/media/77642/download&lr=213&mime=pdf&l10n=ru&sign=5bc08065d038d478209b122441e2ffc4&keyno=0 (accessed on: July 3, 2021).

35. Klump J., Wyborn L., Wu M. et al. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. Data Science Journal. 2021. Vol. 20, No. 12, pp. 1-13.

Best Practices in Medical Imaging

Issue 103

Authors:

Sergey Morozov
Anton Vladzymyrskyy
Anna Andreychenko
Ekaterina Akhmad
Ivan Blokhin
Victor Gombolevsky
Victoria Zinchenko
Nicholas Kulberg
Vladimir Novik
Nikolay Pavlov

DATASET PREPARATION REGULATIONS AND APPROACHES TO REPRESENTATIVE DATA SAMPLING

Part 1

Guidelines

Research Coordination Department, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department

Technical editing by A.I. Ovcharova
Desktop publishing by E.D. Bugaenko

Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies,
Moscow Healthcare Department
24 Petrovka St., Moscow 127051, Russia

+7 (495) 276-04-36 info@npcmr.ru www.tele-med.ai