Instructions to Authors and Reviewers: Medical Physics Dataset Article

A Scope and Submission Guidelines

Medical Physics Dataset Articles (MPDAs) describe scientifically or clinically valuable open-access datasets with high potential for contributing to the research of medical physicists and other investigators working on related problems. In contrast to Research Articles, MPDAs should not include hypothesis testing or data analyses supporting generalizable conclusions. MPDAs should provide detailed descriptions of high impact research/clinical practice datasets, including the value; scope; conditions of acquisition; known limitations; curation and quality assurance processes; and format of the data set. Comprehensive descriptive analysis is required and graphical visualizations are allowed and even encouraged to better understand the dataset. MPDAs should focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses.

As a condition of submission and acceptance of a MPDA, the authors must place the dataset in a recognized and stable data archive that makes the data available to other investigators with minimal restrictions. As with all Medical Physics articles, submitted MPDA manuscripts will be peer-reviewed for possible publication. Editorial evaluation and peer review of MPDAs will consider their novelty, importance to the field, completeness quality of the dataset and accessibility. The data must be made publicly available without restriction in the event that the **Medical Physics Dataset Article** is accepted for publication. MPDAs describing collections of images derived from obsolete imaging devices; images or data acquired under poorly controlled conditions; or datasets lacking sufficient annotations, e.g., physician-drawn contours, segmentation landmark, or clinical diagnoses or clinical outcomes, to support hypothesis-driven research will likely be rejected.

A.1 Data acquisition policies and selection of a repository

It is *Medical Physics'* policy that all key datasets described by a **Medical Physics Dataset Article** manuscript -including computational data, curated data, and data acquired via an experimental or observational procedure– should be placed in an appropriate external repository prior to submission of the manuscript. We believe that this is the best means of making these data discoverable, reproducible and reusable, and we will work with our authors to identify the most appropriate location(s) for their datasets.

Authors should provide their data in the 'rawest' form that will permit substantial reuse. It may be advantageous to release some types of data at multiple levels to enable their broadest reuse, for example, CT sinogram data may best be released as 'raw' readings, including only detector response and background corrections, as well as more processed sinograms including flood fielding and water-equivalent beam hardening corrections. Authors may also submit supplementary information files – including code, models, workflows and summary tables – via the American Institute of Physics' (American Institute of Physics Publishing (AIPP) is our current publisher) Electronic Physics Auxiliary Publication Service (EPAPS). However, primary data should not be submitted as supplementary information unless EPAPS has been approved as the primary repository. All code and information needed for the Referees to access and manipulate the data must be provided at submission. Data, including headers and metadata, derived from patient studies must be fully anonymized and free of protected health information (PHI). All human or animal research studies from which the dataset was derived must have adhered to HHS and all other applicable regulations and ethical guidelines.

Finding appropriate digital repositories for large and complex datasets is challenging. Currently, neither the American Association of Physicists in Medicine (AAPM) nor our publisher are able to provide archiving and curation for large datasets. A trusted archive must

- Be broadly supported and recognized within their scientific community
- Ensure long-term persistence and preservation of datasets (> 10 years) in their published form including backup and web-based hosting capability
- Provide the dataset with a Digital Object Identifier (DOI) that is referenced in the MPDA
- Must be able to support any restrictions on data access required to protect human research subjects. Such restrictions require Editor approval.
- Provide curation that enables datasets to be stored in internationally recognized data formats (preferably DICOM) with appropriately linked metadata

The trusted repository selected by a **Medical Physics Dataset Article** author must be approved by the Editor-in-Chief if it is not on our preapproved list of repositories. The current preapproved repositories include

1. The Cancer Imaging Archive (TCIA) (http://www.cancerimagingarchive.net/)
2. CERN's Zenodo respository. (https://zenodo.org/)
3. Dryad repository (http://datadryad.org/)

Currently, we cannot review MPDA until the dataset has been archived in final form.

Currently our top choice of repository is The Cancer Imaging Archive (TCIA) (http://www.cancerimagingarchive.net/). The TCIA is limited to image-related datasets that have potential for contributing to cancer research. The TCIA welcomes a large variety cancer-imaging relevant datasets, including those useful for radiomic marker development; algorithm validation; and optimization of image-guided therapies. TCIA provides extensive curation resources, including conversion to DICOM format and data anonymization. It is the Editors' hope that the majority of MPDA submissions will be able to utilize TCIA curation for their datasets. TCIA has an internal National Cancer Institute (NCI) steering committee that must independently review and approve all data collections submitted to TCIA. The Editors have met with NCI and have agreed to work together in order to develop a parallel review process that includes both TCIA review and editorial evaluation of the accompanying MPDA manuscript. However, at this time, we cannot review such MPDA manuscripts until the dataset in question has been accepted by TCIA and made available to our Review Team.

The Zenodo and Dryad are uncurated repositories. Neither entity places topical restrictions on the datasets nor do they provide curation services. Quality, integrity, and completeness of the data are the authors' responsibility. Zenodo provides free storage for uploads less than 10 GB while Dryad charges $120 for datasets less than 20 GB. Repository fees are the authors' responsibility.

B Medical Physics Data Article format

MPDAs should be limited to 10 published pages. At the Editor's discretion, additional pages maybe published provided the authors are willing to assume excess page charges.

B.1 Structured Abstract (<300 words)

- **Purpose:** brief summary description of the dataset, including purpose, scope, target audience, and potential applications
- **Acquisition and Validation Methods**: Briefly identify population or phenomenon characterized and data acquisition, processing, and validation procedures
- **Data Format and Usage Notes**: provide data types, number of subjects, population, formats, method of access and link to repository
- **Potential Applications:** Brief description of proposed scientific and/or clinical applications of the dataset and important limitations

B.2 Manuscript structure

B.2.1 **Introduction:**

Summary of the scientific and clinical background of the data set, including a succinct summary of its novelty and expected clinical and/or scientific impact. The use of the data in any prior publications should be described and cited here.

B.2.2 **Acquisition and Validation Methods:**

The Methods should include a detailed description of the experimental or computational procedures used in producing the data, including full descriptions of the experimental design, data acquisition assays, and any computational processing (e.g. normalization, image feature extraction). If the data are derived from observations of animal or human subjects, appropriate regulatory approvals should be cited and the protocol for subject selection and study summarized, including anonymization procedure. Related methods should be grouped under corresponding subheadings where possible, and the methods should be described in enough detail to allow other researchers to interpret and repeat, if required, the full study. Specific data outputs should be explicitly referenced via data citation (see Data Records and Data Citations, below).

For studies using computational tools in the generation or processing of datasets, a statement must be included in the Methods section, under the subheading "computational tools", indicating whether and how the associated code can be accessed, including any restrictions to access. This section should also include information on the versions of any software used, if relevant, and any specific variables or parameters used to generate, test, or process the current dataset. For example, if a condensed history code is used to generate particle tracks, the transport model used must be fully identified, along with all relevant parameter settings (step size, energy cutoffs, etc.) and cross section libraries used. While model and algorithm details maybe referenced, a general overview should be given.

The Data Validation subsection should present any experiments or analyses that are needed to support the technical quality of the dataset. This section may be supported by figures and tables, as needed. *This is a required section*; authors must provide information to justify the reliability of their data.

Possible content **may include:**

- Experiments that support or validate the data-collection procedure(s), e.g. benchmarking of computer codes or instruments against known standards.
- Statistical analyses of experimental error and variation, e.g., reproducibility of mammogram reader scores, measurement of inter-operator segmentation variability, landmark localization error of a nonrigid registration code.
- Consistency checks, e.g., showing that reconstructed CT images from a documented sinogram dataset match vendor-reconstructed images.

- General discussions of any procedures used to ensure reliable and unbiased data production, such as blinding and randomization, sample tracking systems, etc.
- Limitations or uncertainties associated with the dataset
- Any other information needed for assessment of technical rigor by the referees

Generally, this **should not include:**

- Follow-up experiments aimed at testing or supporting an interpretation of the data
- Statistical hypothesis testing (e.g. tests of statistical significance, assessing performance of competing algorithms, trend analysis, etc.)
- Exploratory computational analyses like clustering and annotation enrichment, unless such features are part of the dataset

B.2.3 **Data Format and Usage Notes:**

The Data Format section should open with an overview of the data file structures and their format including the repository where this information is stored and the methodology for accessing the data.

A detailed description should be given of the format of each data record associated with this work and the physical identity and units associated with each data field. For large, complex datasets, tables should be used to succinctly describe data record format and content and should clearly indicate the samples and subjects, their provenance, and the experimental manipulations performed on each with clear references to the methods described in the previous section.

This section should contain brief instructions to assist other researchers with access and manipulation of the data. This may include discussion of software packages that are suitable for analyzing the data, suggested downstream processing steps (e.g. normalization, etc.), or tips for integrating or comparing the data records with other datasets. Authors are highly encouraged to provide code, programs or data-processing workflows if they may help others understand or use the data. Any specialized software tools needed to access the data set or transform it into an accessible data format, such as binary data readers, interpolation code, or recovery of data from linear combinations of basis functions, should be provided in the data repository.

B.2.4 **Discussion**

This section should discuss future applications, analyses, hypotheses; or potential dataset extensions enabled by making the authors' dataset available. Limitations of the proposed dataset with respect to future scientific uses and comparison to competing datasets of the same type should be addressed in this section.

B.2.5 **Conclusion**

A succinct summary description of the dataset, its contribution to the literature, and potential applications

B.2.6 **Acknowledgements**

In addition to acknowledging funding sources and contributions from non-authors, any potential conflicts of interest should be specified

B.2.7 **References**

Follow the formatting requirements outlined in the full Instructions to Authors [LINK]

B.3 Author Checklist

- The expected future utility, e.g., with use cases, is clearly described and illustrated if feasible.
- Quality assurance and validation processes are succinctly described. This would preferentially include a process for examining the data at some specified level of detail, using software or processes that are widely available to any medical physicist, or specialized software made available with the data.

- The use of any standards should be clearly described. It is highly desirable that existing ontologies and other standards are used where available.

- The manuscript should include a table that lists the key data types, key metadata, and the fraction of data for which each element is available. Preferably 100% of all key data fields are available, but this is often not possible with clinical data.

- The process for locating the data and accessing the data is clearly described. This is through a very well established repository: posting on an institutional or laboratory website is generally not suitable. The dataset and associated programs must be identified via a 'DOI' (digital object identifier), DOIs are available through repositories.

- Data should be completely freely available. Exceptions will be granted only in cases where privacy, e.g., protection of personal health information (PHI), or public safety is a concern. For such cases, user registration and/or execution of a data sharing agreement maybe required as a condition of access..

- Note that a single MPDA could describe several related data sets that are useful within the same type of investigation.

- The adoption of any standards for nomenclature, semantic interoperability, or other ontology's should be well described.

- The **Acquisition and Validation Methods** section shall include detailed step-by-step description of procedures used to generate or acquire the data the data. Any publications that have previously cited the data should be clearly referenced. There should be enough detail for users to potentially add new data that is consistent with the submitted data set.

- The **Data Format and Usage Notes** section should give practical guidance for accessing and interpreting the data set.

- If software systems were used to generate the data, they should be described in full detail, including version numbers. Any deviation from this should be justified. If the software is not generically available, and it has an open source license, it should be archived with the data set.

- Any filtering or other selection process used to select and generate the final data set should be clearly described, potentially in a flowchart diagram.

- The **Discussion** section should include qualitative comments on any underlying drawbacks or advantages to the data that may not be obvious.

- Data sets must be de-identified. Upon request, the MPDA Review Team can provide specific guidance on tools to use, handling of the many fields in DICOM files where PHI can lurk, approaches for handling of dates/delta dates, age ranges, etc.

- MPD authorship guidelines:
    - All authors must have read and approved submission of the final version of the manuscript
    - Each author must have contributed substantially to the data collection process, the design of the data collection process, or to the data stewardship processes, e.g., quality assurance or creation of software tools access the data.
    - Providing funding or resources, clinical access to patient subjects/information, or purely technical contributions alone do not qualify one for coauthorship

B.4 Instructions to Referees:

1. The Review Team must include an expert in the research domain addressed by the dataset. This referee should address the following issues
    a. dataset novelty
    b. dataset relevance and importance to addressing significant research questions in this domain
    c. Adequacy and quality of the dataset is sufficient to contribute to targeted research applications
2. The Review Team must include an expert in dataset access and curation. This referee should address the following issues
    a. verify existence/accessibility of dataset via supplied DOI or other link
    b. verify that data can be successfully downloaded
    c. verify the software tools supplied or referenced can open dataset
    d. , verify number and type of image sets or data structures
    e. graphically/qualitatively evaluate the accuracy/integrity/consistency of one randomly selected data structure
    f. If a repository has already performed a curation process on the dataset (e.g., the TCIA) this should be noted.
    g. Access and analyze downloaded data as needed to verify fitness for the envisioned purpose.