Mass General Brigham

Guidance for Sharing Model Code and Data as a Requirement for Publication

Latest Revision: 08.1.22

Purpose: As AI/ML research continues to expand, there are increasing privacy concerns stemming from the publication or dissemination of model code, hyperparameters, weights, and training datasets. At the same time, in an effort to advance the field, ensure scientific reproducibility, and embrace open science, many journals and granting agencies are requiring machine learning model code with or without associated data to be shared as a requirement for publication. This document provides guidance for researchers as they consider when to publish and what elements to share as part of a discrete individual research project (i.e. not part of an on-going open source activity).

Table of Contents:

A. Definitions

1. **Model Data Input Types:** training data used in the derivation of model weights includes, but is not limited to: structured and unstructured text, numeric data, pixel/image data, categorical data, time series data, biomedical signal data (e.g., EKG, EEG, wearables), genomic data, voice and audio data, and dataset labels and annotations.

2. **Model Architecture with Hyperparameters:** the model architecture or topology includes the learning algorithm or method (e.g., Convolutional Neural Network, Random Forest, Support Vector Machine), the model design (e.g., Activation Function, Loss Function, Optimizer, size of network / number of layers and number of units per layer, weight initialization, Dropout Layer), and the set of features or hyperparameters that affect the speed and quality of the learning process (e.g., Learning Rate, Dropout, Batch Size).

3. **Model Code:** The source or computer code that is a human-readable set of instructions, functions, libraries, and statements written in a programming language (e.g., Python, C++) that is executable by an interpreter or can be compiled into an executable file. Model code includes source code as well as the accompanying non-coding comment lines. This definition of model code is inclusive of the preprocessing steps (e.g. coded data manipulations or calculations).

Further, we distinguish between the following functionality that the source code may implement:

- **Training code:** The source code required to produce a set of trained model weights given raw input data. Typically, this will include tasks such as data preprocessing, an implementation of the model architecture, and the optimization procedure used to train the model.
- **Inference code:** The source code required to run the model on novel raw input data to produce the desired model output. This will typically include tasks such preprocessing, execution of the model, and post-processing of the model output.

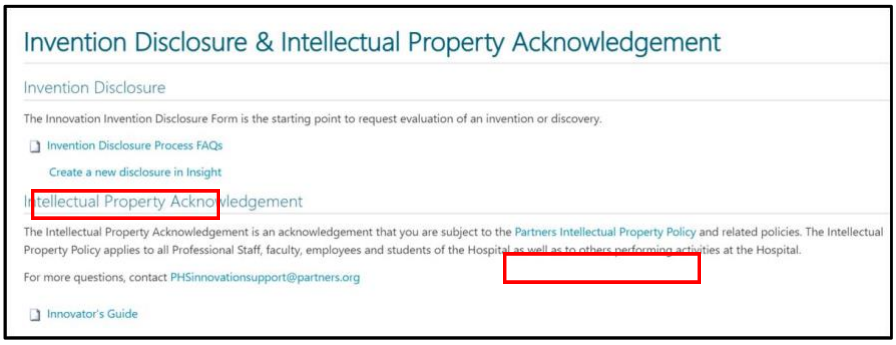
Depending on the situation, it may be desirable to release either the training code or the inference code, or both. Generally speaking, in the context of publication it would not make sense to release the inference code without releasing either the model weights or the training code.

4. **Model Weights:** parameter values learned/derived via the training process that control the signal strength of the connection between nodes/layers of the model, deciding the amount of influence a specific input/feature will have on the output.

5. **Model:** is the model architecture, the source code for inference, and the model weights.

6. **Shared Assets:** Model Architecture with Hyperparameters, Executable File or Environment implementing the model (e.g., Docker), Training and/or Inference Source Code, Model Weights, and/or Training Data. This definition applies only in the instance of a discrete individual research project, not part of an on-going open-source activity.

7. **Software[1]:** Means computer or computer-based materials in the broadest sense, including but not limited to computer programs, user interfaces, users' manuals and other accompanying explanatory materials or documentation, mask works, firmware, and computerized databases. For convenience, all Software developed by MGB faculty, trainees and staff is treated as Copyrightable Work for purposes of the MGB Intellectual Property Policy; in many cases, Software materials will constitute Inventions and will need to be disclosed to MGB Innovation via the Invention Disclosure Form.



[1] This definition is consistent with the Intellectual Property Policy for Partners Affiliated Hospitals and Institutions

Mass General Brigham

B. Guidance for Use of Image De-identification Services

I. De-identification of Text and Pixel Data:

Healthcare data protection Privacy and Security Rules embedded within the Health Insurance Portability and Accountability Act (HIPAA) of 1996, were augmented by the Privacy Rule of 2003 that codified the concepts of Protected Health Information (PHI). Often used interchangeably, Anonymization and De-identification are distinct processes for removal of PHI from data sets to protect the privacy of patients' personal health information and hinder the possibility of tying data back to an individual patient.

- **Anonymization** removes PHI from a data set and destroys any link to the patient's original identity; whereas
- **De-identification** requires the removal or replacement of data corresponding to the 18 categories of PHI defined by HIPAA[2,3]. Often, the de-identification process replaces an actual patient identifier with a pseudonymized label. De-identified data may be coded with the creation of a mapping key between the actual and pseudonymous identity of the patient. This link to the original, fully identified data set may be kept by an honest broker. Links exist in coded de-identified data making the data considered indirectly identifiable and therefore not anonymized.

It's important to recognize the potential locations of PHI when attempting to de-identify or anonymize medical imaging data files:

- **DICOM Header Metadata:** DICOM is the primary file format for storing and transferring medical images; each file contains a header that includes PHI in prescribed file locations.
- **Vendor Specific Private Attributes:** Some image acquisition vendors use private attributes placed in optional shadow fields in the DICOM header, the removal of which may hinder processing of the images required for the research study. *A manual check of these shadow fields should be done to ensure no PHI is written in them.*
- **Burned-in Pixel Data:** Some modalities (e.g., ultrasound and 3D reconstructed or advanced post-processed CT and MR images) burn-in PHI into the pixel data. De-identification of

---

[2] US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the HIPAA Privacy Rule. Available at: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed January 31, 2022.

[3] HIPAA Identifiers: 1. Names; 2. All geographical identifiers smaller than a state, except for the initial three digits of a zip code if, according to the current publicly available data from the U.S. Bureau of the Census: the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; the initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000; 3. Dates (other than year) directly related to an individual; 4. Phone numbers; 5. Fax numbers; 6. Email addresses; 7. Social Security Numbers; 8. Medical record numbers; 9. Health insurance beneficiary numbers; 10.Account numbers; 11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers; 13. Device identifiers and serial numbers; 14. Web Uniform Resource Locators (URLs); 15. Internet Protocol (IP) address numbers; 16. Biometric identifiers, including finger, retinal and voice prints; 17. Full face photographic images and any comparable images; 18. Any other unique identifying number, characteristic, or code except the unique code assigned by the investigator to code the data

burned-in pixel data requires spatial location of the PHI in the image and blackening of those pixels. *It is recommended that a manual review of a random sample of images for each vendor make and model should be done to determine whether there is PHI on the images.*

- **Facial Characteristics:** Defacing of an image set may be needed to reduce the potential for re-identification of a patient through facial characteristics from a 3D reconstructed imaging examination. Several methods exist including masking, distortion of features and/or removal of regions in the image containing the eyes, nose, and mouth.

Regardless of the tool used, the PI is ultimately responsible for ensuring the dataset is accessed, stored, and managed in a way that is consistent with an approved IRB protocol, including de-identification if required.

We strongly recommend that the PI conducts a manual audit post- de-identification to ensure all PHI and other potential patient identifiers (e.g. unique markings or anatomy) have been removed. or masked in accordance with the approved IRB protocol.

II. The MIDAS De-identification Tool:

The **Medical Imaging Data As A Service (MIDAS)**, is a suite of self-service tools that enable MGB investigators to obtain radiologic imaging studies for research purposes and AI/ML development. As part of this service, there is an option for images to be de-identified. This tool modifies the relevant DICOM header fields in order to comply with HIPAA de-identification regulations and MGB IRB requirements. The tool works by either deleting or transforming the DICOM information; please refer to the Compass Default Tags document for the specific permutations applied to each field. For more information about specifications for the DICOM fields that will be automatically de-identified or how to request customization, please follow the instructions provided here.

Please note the automated de-identification tool available for use with MIDAS does **NOT**:

- remove **Vendor Specific Private Attributes**.
- remove **Burned-in Pixel Data**.
- deface or remove **Facial Characteristics**.

III. Additional De-identification Tools/Processes:

In addition to the MIDAS de-identification tool, there are other tools which are freely or commercially available for de-identification of DICOM data, burned-in pixel data, or defacing. They may require substantial customization for adequate removal of PHI[4] and often manual quality reviews of the process are needed. Known tools are listed below:

- RSNA Clinical Trials Processor (CTP)
    - Aides with the removal of Burned-in Pixel Data

[4] Batlle JC, Dreyer KJ, Allen B, et al. Data Sharing of Imaging in an Evolving Health Care World: Report of the ACR Data Sharing Workgroup, Part 1: Data Ethics of Privacy, Consent, and Anonymization. *J Am Coll Radiol* 2021;18:1646-1654.

    - Allows specification of specific metadata fields to keep during the de-identification processes that otherwise removes DICOM fields
- DICOM Library
- GDCM
- PixelMed DICOMCleaner
- Tudordicom
- YAKAMI DICOM
- ACR's Transfer of Images and Data, ACR Connect
- SynthStrip (skull stripping)
- MiDeFace (Defacing images)
- Many commercial PACS include de-identification tools

Mass General Brigham

C. Decision Tree

I. Preface:

Journals may request that you share your machine learning model code and/or data for the work that is covered in your manuscript and has been accepted for publication. The following sections walk through **two decision trees: (1) Sharing Data and (2) Sharing Model Code**. Each decision tree contemplates the following and will impact **whether** and **what** you share:

1. Your **desire to commercialize** your model;

   If you have **questions about commercialization** and the value of your model, please review MGB's Intellectual Property Policy for Partners Affiliated Hospitals and Institutions Section B.4 and Section B.6, and the Software Licensing Guidance for PIs and the Intellectual Property Policy Section C.11.1 'Works to be Disclosed'. If necessary, fill out an Invention Disclosure Form on the Insight System. You may reach out to the MGB Innovation Office for assistance.

2. Your **funding source** (e.g., a commercial funding source may preclude this); and/or
3. Your current **approved IRB.** If your approved IRB does not contemplate sharing and broad dissemination, you must amend your IRB to allow sharing of de-identified data.

Other Considerations:

Risk/Value of the Shared Assets: You may be able to negotiate what **Shared Assets** are included for the publication depending on the journal requests and your individual circumstances. The **Shared Assets** have different associated risks and value (in order of increasing risk/value):

- Model Architecture with Hyperparameters
- Executable File or Environment implementing the model (e.g., Docker)
- Training and/or Inference Source Code
- Model Weights
- Training Data

Data Use Agreements: A Data Use Agreement (DUA) may be required by your funder, MGB, or the dissemination platform you select.

Mass General Brigham

II. Decision Tree: Sharing Data

Please note: The MGB Intellectual Property Policy for Partners Affiliated Hospitals and Institutions **requires** the Invention Disclosure Form be completed via Insight if you are making the Work available to the public or any third party and *either* are reasonably likely to use it for commercial purposes.

1. Commercialization

1. Do you **plan to commercialize** your model and/or is there potential intellectual property value in the model?
   - **NO**: You may comply with the journal's request
     - The data's level of identifiable risk (e.g. de-identified, limited data set (LDS), PHI) needs to be consistent with your MGB IRB approved protocol for sharing.
     - A DUA may be required.
   - **YES:** Request exemption from the journal to not share the data
     - **IF DENIED EXEMPTION**: Can you share a subset of the data?
       - **YES:** You may comply with the journal's request
         - The data's level of identifiable risk (e.g. de-identified, limited data set (LDS), PHI) needs to be consistent with your MGB IRB approved protocol for sharing.
         - A DUA may be required.
       - **NO:** If request to share a subset of the data is denied and the journal requires sharing of all training data, consider delaying publication until commercialization activities have concluded.

2. Funding Source

2. Does your **Funding Source** allow and/or require sharing of the data?
   - **YES:** You may comply with the journal's request
     - The data's level of identifiable risk (e.g. de-identified, limited data set (LDS), PHI) needs to be consistent with your MGB IRB approved protocol for sharing.
     - A DUA may be required.
     - The sharing is consistent with any funding source restrictions.
   - **NO:** Request exemption of sharing the data due to funding source restrictions

3. Approved IRB / Institution

3. Does your MGB **IRB/Institution** (e.g., Data and Tissue Sharing Committee) allow sharing of the data?
   - **YES:** You may comply with the journal's request
     - The data's level of identifiable risk (e.g. de-identified, limited data set (LDS), PHI) needs to be consistent with your MGB IRB approved protocol for sharing.
     - A DUA may be required.
   - **YES, WITH RESTRICTIONS:** You may share data if they meet the stipulations set forth by your approved IRB or by the MGB Data & Tissue Sharing Committee
   - **NO:** If special or restricted dataset (e.g., neonatal brain MRIs with rare pathology)
     - Request exemption of sharing data

Mass General Brigham

III. Decision Tree: Sharing Model Code

Please note: The MGB Intellectual Property Policy for Partners Affiliated Hospitals and Institutions **requires** the Invention Disclosure Form be completed via Insight if you are making the Work available to the public or any third party and *either* are reasonably likely to use it for commercial purposes.

1. Commercialization

1. Do you **plan to commercialize** your model code and/or is there potential intellectual property value in the model?
   - **NO:** You may comply with journal's request
     - Attach a copyright / license (e.g., BSD3) as a disclaimer against liabilities and to establish terms of use. For examples, see *Section D. Commonly Used Licenses*.
   - **YES:** As above, submit MGB Innovation's Invention Disclosure Form via Insight for your model code and determine next steps consistent with MGB's Intellectual Property Policy for Partners Affiliated Hospitals and Institutions.
     - For consideration and in evaluation with MGB Innovation, alternatives to sharing the model include requesting an exemption from the journal, sharing of the training Model Code only (without model weights; not the inference code), or delaying publication until commercialization activities have concluded.

If you have **questions about commercialization** and the value of your model, please review MGB's Intellectual Property Policy for Partners Affiliated Hospitals and Institutions Section B.4 and Section B.6, and the Software Licensing Guidance for PIs and the Intellectual Property Policy Section C.11.1 'Works to be Disclosed'. You may reach out to the MGB Innovation Office for assistance.

2. Funding Source

2a. Does your funding source **require** sharing of model code (e.g., NIH)?
- **YES, FUNDING SOURCE REQUIRES SHARING:** You must comply with funder's requirements and may comply with the journal's request.
  - Attach a copyright / license (e.g., BSD3) as a disclaimer against liabilities and to establish terms of use. For examples, see *Section D. Commonly Used Licenses*.

2b. If your funding source does not require sharing, do they **allow** sharing of model code?
(Assess if the funder's allowance for sharing has restrictions or stipulations. These restrictions or stipulations may or may not align with the journal's request).
- **YES, FUNDING SOURCE ALLOWS SHARING:** You may comply with journal's request.
  - Attach a copyright / license (e.g., BSD3) as a disclaimer against liabilities and to establish terms of use. For examples, see *Section D. Commonly Used Licenses*.
- **NO, FUNDING SOURCE DOES LIMITS SHARING:** (e.g. model code is considered proprietary) Request exemption from the journal to not share the model code.
  - **IF EXEMPTION REQUEST IS DENIED:** Consider alternatives to sharing the model code such as sharing the training Model Code only (without model weights; not the inference code) or **delaying publication until commercialization activities have concluded**.

Mass General Brigham

D. Commonly Used Licenses

It is recommended that a copyright / license be attached as a disclaimer against liabilities and to establish terms of use. The following are examples of commonly used licenses.

I. Software Licenses

Ia. "Permissive" Software Licenses

- Allow users to do most things with the code provided attribution is given
- Examples
    - MIT License (https://en.wikipedia.org/wiki/MIT_License)
    - Apache License (https://en.wikipedia.org/wiki/Apache_License)
    - BSD License (https://en.wikipedia.org/wiki/BSD_licenses)
    - The 3D Slicer license is a modified BSD style license (3D Slicer license)
    - LGPL (Lesser GNU General Public License; https://en.wikipedia.org/wiki/GNU_Lesser_General_Public_License). This is free for use, but requires extension of the code to be shared.

Ib. "Copyleft" Software Licenses

- Require any redistributed version to be made available under the same terms
- In practice, this means that it is difficult for others to use code distributed under a copyleft license in a commercial product (but it does not preclude any company using the code internally for business purposes – redistribution is the key here).
- Examples
    - GPL (GNU General Public License; https://en.wikipedia.org/wiki/GNU_General_Public_License)
    - Mozilla License with Healthcare Disclaimer
        - Open software license with associated rider language that helps limit institutional liability. Under development.

II. Creative Commons License (https://creativecommons.org)

Purpose: for "non-code" assets, such as datasets and trained model weights.

- Subtypes of Creative Common Licenses: https://creativecommons.org/licenses

III. General Resource

License Selection Guidance: https://choosealicense.com (created by Github)

Mass General Brigham

E. Commonly Used Sharing Platforms*

The following is a non-exhaustive compilation of platforms that may be appropriate for sharing your model code and/or data. **Please review the specific requirements before using any platform for sharing.*

I. Github (https://github.com)

- Go-to for both developers and scientists. *De facto* standard.
- Common platform for sharing code (not as well suited to shared data)

II. Gitlab (https://about.gitlab.com)

- Common alternative to Github with a near identical feature list and can be self-hosted
- MGB currently uses Gitlab internally (https://gitlab.partners.org within MGB firewall; not accessible to the general public)
- Common platform for sharing code (not as well suited to shared data)

IV. Open Science Framework (https://www.cos.io/products/osf; https://datamanagement.hms.harvard.edu/share/data-repositories/open-science-framework)

- OSF integrates with Github for code sharing

V. Other Platforms - Harvard Data Sharing Platforms Matrix (https://docs.google.com/spreadsheets/d/1_VbZsxYeov1Z19Cf5CkwToDW1WpWLLCS9ZKG7aCmJGc/edit#gid=0) A list of Data Sharing platforms has been compiled by the Harvard Longwood Medical Area Research Data Management Working Group which is not formally affiliated with MGB. Please note, these platforms have not been vetted and are provided as additional resources; you should review specific requirements before using any platform for data sharing.

Mass General Brigham

Harvard Longwood Medical Area Research Data Management Working Group	✓ Yes ✗ No	Page last updated November 20, 2020						
Requirement	**Dataverse**	**Dryad**	**figshare**	**GigaScience**	**Mendeley Data**	**OSF**	**Vivli**	**Zenodo**
Data Size and Format								
Hosting of common file formats (e.g. csv, tsv, xls, xlsx, doc, pdf)	✓	✓	✓	✓	✓	✓	✓	✓
Hosting of proprietary file formats (e.g. raw image files)	✓	✓	✓	✗	✓	✓	✓	✓
Unlimited size per file	✗	✓	✗	✓	✗	✗	✓	✗
Unlimited total dataset size	✓	✓	✓	✓	✓	✓	✓	✓
Data Licensing								
CC0 waiver	recommended	required	recommended	required	available	available	available	available
Data Attribution and Citation Tools								
Assignment of dataset DOIs	✓	✓	✓	✓	✓	✓	✓	✓
User Access Controls								
Tiered access (e.g. administrator-level, collaborator-level, curator-level)	✓	✗	✓	✗	✓	✓	✓	✗
Journal-integrated, anonymous access (for peer review pre-publication)	✓	✓	✓	✓	✓	✓	✗	✗
Optional embargo to data release following publication	✗	✓	✓	✓	✓	✓	✓	✓
Data Access Tools								
Comprehensive data and metadata search tools	✓	✗	✗	✗	✓	✓	✓	✗
Data access via direct download	✓	✓	✓	✓	✓	✓	✓	✓
Data downloading via API	✓	✓	✓	✗	✓	✓	✗	✓
Built-in tools for reading proprietary file formats	✗	✗	✓	✗	✗	✗	✓	✗
Integrated data analysis tools	✓	✗	✗	✓	✗	✗	✓	✗
Cost								
Data deposition fees	none	tiered	none	none	none	none	membership	none
Data maintenance fees	none	none	none	none	none	none	membership	none