Data, Licensing & Ethics 101

A Practical Guide for Open Data MICCAI 2026 Authors

This document is intended for anyone submitting a dataset and paper in Open Data MICCAI 2026. This guide explains the concepts, requirements, and practical steps behind the data, licensing and ethical considerations sections of your submission.

You do not need a legal or ethics background to follow this guide. Where decisions require institutional input, we tell you clearly.

Part 1: What Makes Your Data "Open"?

The basic idea

Open data does not simply mean data that is accessible online. A file you can download if you email the right person is not open data.

Open data means that anyone, anywhere, at any time, can find, access, and use your data without needing to ask permission, sign a bilateral agreement, or belong to a specific institution.

This distinction matters because a significant proportion of datasets described as "publicly available" in published papers are, in practice, inaccessible. Links break. Institutional servers go offline. Corresponding authors change jobs. The data effectively disappears. Open Data MICCAI exists precisely to raise the bar: we require that datasets are genuinely, sustainably, and legally open at the time of submission and beyond.

Part 2: Licensing

What is a license?

A license is a legal instrument that tells others what they are, and are not, allowed to do with your dataset. Without a license, your data is "all rights reserved" by default under copyright law in most countries. This means that even if your data is publicly downloadable, nobody can legally reuse, redistribute, modify, or incorporate it into their own work without your explicit written permission.

For a dataset to be open in any meaningful sense, it must have a license attached to it. The license is what converts "technically accessible" into "legally usable." A license is not the same as a data use agreement or any other data agreement. A license is a one-way declaration you attach to your dataset; users are automatically bound by its terms simply by using the data, with no paperwork required.

Creative Commons licenses

For datasets, different from software, the Creative Commons (CC) family of licenses is the international standard for open data sharing. CC licenses are free, legally robust, internationally recognised, and accepted by all major repositories, journals, and funders.

Every CC license requires attribution (users must credit you). Beyond that, they vary on two axes:

Commercial use: Can users use your data for commercial purposes?

- If yes: use a standard CC license (e.g. CC BY)
- If no: add the NC (Non-Commercial) modifier (e.g. CC BY-NC)

Derivative works: Can users modify, preprocess, or build upon your data and share the result?

- If yes, freely: use a standard CC license
- If yes, but only under the same terms: add the SA (Share-Alike) modifier (e.g. CC BY-SA)
- If no: add the ND (No Derivatives) modifier — but see the warning below

The resulting options relevant to Open Data MICCAI are:

License	Commercial use	Derivatives	Our assessment
CC BY 4.0	✅ Yes	✅ Yes	⭐ Recommended. Maximum reuse. Accepted by all major funders.
CC BY-SA 4.0	✅ Yes	✅ Yes (same terms)	✅ Acceptable. Derivatives must remain open.
CC BY-NC 4.0	❌ No	✅ Yes	✅ Acceptable if your ethics or institution requires NC restriction. Explain why in your paper.
CC BY-NC-SA 4.0	❌ No	✅ Yes (same terms)	✅ Acceptable with the same caveat as above.

CC BY-ND / CC BY-NC-ND	—	❌ No	⁉️ Acceptable with the same caveat as above but not recommended.

Why CC BY 4.0 is our recommendation.

It maximises the scientific value of your dataset. It is the default requirement of major funders including the NIH, Wellcome Trust, and European Research Council. It is accepted without exception by Zenodo, PhysioNet, TCIA, and most other repositories. It allows other researchers to preprocess, annotate, augment, and incorporate your data, which is precisely the kind of secondary use that advances medical AI.

Why CC BY-ND is not recommended. The ND (No Derivatives) clause prohibits creating modified versions of the data. In medical imaging research, many downstream uses involve some form of modification: format conversion, intensity normalisation, resampling, combining with other datasets. CC BY-ND would prohibit all of this, making the dataset legally unusable for standard research workflows. Submitting under CC BY-ND is therefore inconsistent with the open data mission of this initiative.

How to choose your license in under two minutes

The Creative Commons License Chooser tool walks you through the decision with two yes/no questions:

👉 https://creativecommons.org/chooser/

1. Open the link. You will see two questions.
2. **"Allow adaptations of your work to be shared?"**
    a. For most datasets, answer Yes. If you want to ensure modified versions always remain open under the same terms, choose Yes, as long as others share alike.
3. **"Allow commercial uses of your work?"**
    a. If your ethics approval or institutional policy does not restrict this, answer Yes. If you have a documented restriction, answer No.
4. The tool will display your license with a badge, a human-readable summary, and a permanent URL. Copy the license name (e.g. CC BY 4.0) and URL (e.g. https://creativecommons.org/licenses/by/4.0/) into your repository and your paper.

Licensing for accompanying code and software

CC licenses are designed for creative and data works, not for software. If you are releasing code alongside your dataset, e.g., preprocessing pipelines, annotation tools, model training scripts, that code requires a separate software license. We recommend:

- MIT License — maximum permissiveness, minimal requirements

- Apache 2.0 — similar to MIT, with an explicit patent grant

Both are compatible with CC BY 4.0 data licenses and are accepted by all major code repositories. Include a separate LICENSE file in your code repository.

Upstream license compatibility

If your dataset is derived from or builds upon previously released data, your chosen license must be legally compatible with the original. This is not always obvious. Common scenarios:

- Original data: CC BY 4.0 → Your dataset: CC BY 4.0 ✅ Compatible
- Original data: CC BY-NC 4.0 → Your dataset: CC BY 4.0 ❌ Not compatible, you cannot drop the NC restriction
- Original data: CC BY-SA 4.0 → Your dataset: CC BY 4.0 ❌ Not compatible, you must carry forward the SA requirement
- Original data: custom institutional license → Your dataset: CC BY ⚠️ Check the original terms carefully before assuming compatibility

If you are unsure whether your license is compatible with the source data, contact your institution's legal or data governance office before submission.

Do not assume compatibility. Downstream license conflicts are one of the most common causes of post-publication retraction.

What a license does NOT cover

A license governs intellectual property, the legal right to copy, share, and modify. It does not:

- Replace ethics approval
- Guarantee that sharing is lawful under data protection regulations (GDPR, HIPAA, etc.)
- Prohibit misuse in the ethical sense (a CC license technically permits someone to use your data to train a discriminatory model)
- Constitute informed consent or replace it

This is why licensing requirements must be read alongside the ethics requirements below.

Part 3: Anonymisation and What It Actually Means

The goal and the gap

The goal of anonymisation is to ensure that no individual whose data appears in your dataset can be identified – directly or indirectly – by anyone accessing the data. This sounds straightforward but is significantly more complex in medical imaging than in other domains,

because medical images contain identifying information in multiple layers that are not always obvious.
Anonymisation for open data is not the same as de-identification for internal research use. The threshold is higher because the data will be available to anyone, indefinitely, and you cannot control who accesses it or what other datasets they might combine it with.

Layers of identifying information in medical imaging data

Direct identifiers in metadata. DICOM files contain structured metadata fields including patient name, date of birth, patient ID, institution name, acquisition date, and device serial numbers. These must be removed or replaced. Standard tools for DICOM anonymisation include pydicom, DicomCleaner, and CTP. Simply removing the fields is not always sufficient – some fields contain identifiers embedded in free-text that require parsing (e.g. study description fields that include patient initials).

Indirect identifiers in metadata. Even after direct identifiers are removed, the combination of acquisition date, institution, and clinical variables can be sufficient to re-identify a patient, particularly for rare conditions. Dates should be shifted (not zeroed) using a consistent offset per patient. Institution names should be removed or replaced with a site code.

Identifiers in image data. Some modalities carry identifying information within the image itself:

- MRI head scans may allow facial reconstruction from 3D volumes. De-facing (removing the facial surface from T1 volumes) is required for head MRI datasets. Standard tools include pydeface and mri_deface.
- Fundus and OCT images may contain visible vessel patterns that are unique to individuals (retinal biometrics).
- Pathology slides may contain patient identifiers burned into the slide image during scanning.
- Endoscopic or surgical video may capture faces, operating theatre identifiers, or staff present in the room.

Identifiers in annotation data. Free-text clinical notes, radiology reports, or annotation comments may contain patient names, dates, or other identifying information. All free-text fields must be reviewed and cleaned before release.

What anonymisation does not mean

Anonymisation does not mean:

- Removing only the patient name field in DICOM
- Removing only names
- Converting to a different file format without addressing the DICOM headers first

- Assuming that because images "look anonymous" they are legally anonymous

Pseudonymisation vs anonymisation

These are not the same thing, and the distinction has legal consequences under GDPR at least in Europe:

- Pseudonymisation replaces identifiers with a code or pseudonym, while retaining a mapping table that allows re-identification. Pseudonymised data is still considered personal data under GDPR, because re-identification is possible. It is appropriate for internal research use but is not sufficient for open data release.
- Anonymisation is an irreversible process. No mapping table is retained, and re-identification is not reasonably possible even by the data controller. Only truly anonymised data falls outside the scope of regulations and can be released as open data without ongoing data protection obligations.

Your paper must state explicitly which of these applies and describe the procedures used to achieve it.

Anonymisation pipelines

Where possible, your anonymisation pipeline should be:

- Documented – describe every step in the Methods section
- Reproducible – provide the code used, archived in a permanent repository
- Validated – describe how you verified that anonymisation was complete (e.g. manual review of a random sample, automated scanning for residual identifiers)

If you used a commercial or institutional anonymisation tool, name the tool, version, and configuration settings.

Part 4: Ethics for Open Data Release

Ethics approval is necessary but not sufficient. Ethics approval authorises you to collect and use data within the scope of your study. It does not automatically authorise you to release that data openly. Open data release is a separate step that requires:

1. Your ethics approval explicitly covering secondary use and open release, or a separate amendment approving it
2. Informed consent from participants that covers this use, or a formal consent waiver
3. Completion of anonymisation to the standard required for open release
4. Confirmation that no third-party rights (e.g. vendor contracts, collaborator agreements) prohibit redistribution

Do not assume that because your data collection was ethically approved, open release is permitted. Check your original approval documentation carefully, and if in doubt, submit a formal query to your ethics committee.

Special categories of data

Medical imaging data is almost always "special category" personal data under GDPR (and its equivalents in other jurisdictions) because it relates to health. This means it carries the highest level of data protection obligations, even after pseudonymisation. True anonymisation removes data from the scope of GDPR and other regulations entirely. Until that standard is met, standard data protection obligations apply.

Additional care is required for:

- Paediatric data – children cannot provide legally valid consent in most jurisdictions; consent must come from a parent or guardian, and specific protections may apply
- Data from individuals who were incapacitated at the time of collection – similar considerations apply
- Psychiatric or neurological data – may carry additional stigma risk and require heightened anonymisation
- Genetic or genomic data – considered separately under many regulatory frameworks and may require additional review

The right to withdraw

Under GDPR and many other frameworks, individuals retain the right to withdraw consent and request deletion of their data. For a dataset that has already been openly released, this right creates a practical tension: once data is downloaded and incorporated into third-party datasets or trained models, it cannot be recalled.

Your paper must address this by documenting one of the following:

- The legal basis for data processing was not consent (e.g. public interest under GDPR), in which case withdrawal rights do not apply in the same way
- Participants were explicitly informed at consent that, once released, open data cannot be recalled, and they accepted this
- A mechanism exists (e.g. a flagged version of the dataset) to remove identified records if a withdrawal request is received, and you commit to releasing updated versions in that event

Intended use and misuse risks

Your paper must include a statement of intended use – what tasks, applications, and research questions the dataset is designed to support – and a statement of known misuse risks. For medical imaging data intended for AI development, this should explicitly consider:

- Demographic bias: Is the dataset representative of the population it is intended to model? Known underrepresentation of age groups, sexes, ethnicities, or disease severity levels should be documented
- Discriminatory model training: Could the dataset be used to train models that perform systematically worse for specific subgroups? What is known about this?
- Re-identification risk: Even after anonymisation, what is the residual risk, and for which subgroups is it highest?
- Deployment in unvalidated settings: The dataset should not be used to develop or validate models for clinical deployment without additional validation in the target population and setting

Ethical use clauses

A CC license permits any legal use of your data, including uses you might consider ethically problematic. If you wish to place ethical conditions on use – such as prohibiting re-identification attempts, requiring users to report suspected data breaches, or restricting use to non-surveillance applications – an ethical use clause provides a mechanism to do so.

We recommend considering the Responsible AI Licenses (RAIL) framework (https://www.licenses.ai/), which provides ready-made license templates designed specifically for AI and ML datasets and models. RAIL licenses embed ethical use conditions directly into the license, making them legally binding rather than advisory. If you do not attach an ethical use clause, your paper should explain why the chosen license is considered sufficient given the nature of the data and its intended use.

Part 5: Multi-Site and Multi-Centre Data

Datasets collected across multiple institutions introduce additional complexity in three areas: ethics, data governance, and technical harmonisation.

Ethics approval in multi-site studies

Each contributing institution may require its own ethics approval or a formal data sharing agreement authorising transfer of data to the coordinating site. A single ethics approval from the coordinating institution does not automatically cover data collected at partner sites in most jurisdictions.

Before submission, confirm that:

- Each contributing site has provided ethics approval or a formal exemption statement
- Each site's approval explicitly covers the secondary use and open release of data
- If data was transferred between institutions or countries, the legal basis for that transfer is documented (particularly relevant for transfers of EU personal data under GDPR)

Your paper must list the ethics approval number for each contributing site, or explain the legal basis for exemption.

Informed consent in multi-site studies

Consent forms may vary across sites. A participant at one site may have consented to broader secondary use than a participant at another. Before releasing a multi-site dataset, ensure that the consent obtained at all contributing sites is compatible with open release. If any site's consent does not cover open release, data from that site must either be excluded from the open release or a consent waiver must be sought from that site's ethics committee.

Site harmonisation

Differences in scanner manufacturer, field strength, acquisition protocol, and image reconstruction across sites introduce systematic variability that users of your dataset need to understand. Your paper must:

- Document acquisition parameters per site (not just aggregate)
- Describe any harmonisation steps applied (e.g. intensity normalisation, ComBat harmonisation)
- Provide site labels in the released metadata so that users can account for site effects in their analyses
- Describe known residual site effects and their potential impact on downstream model performance

Where to Get Help

If you are unsure about any aspect of this guide, the following resources are available:

- Licensing questions: Use the CC License Chooser at https://creativecommons.org/chooser/.
- Responsible AI Licenses: https://www.licenses.ai/.
- For complex upstream compatibility questions, consult your institution's legal or data governance office.
- Ethics questions: Contact your institutional ethics committee or IRB.
- For questions specific to Open Data MICCAI requirements, contact the committee at: laura.arbelaez@ub.edu