AIDA Data Sharing Policy

Version: 1.0
Approved: 2019-12-13

Author: Joel Hedlund, Data director AIDA.

Preface

This document has been drawn up by AIDA in order to provide succinct and understandable guidelines for its activities, and in an attempt to describe the common practice in ethical, legal and FAIR data sharing for AI in medical imaging diagnostics in Sweden.

AIDA policies are set continuously by AIDA management, who reports to the AIDA steering group.

AIDA is a collaboration arena for AI in medical image analysis, and is an initiative within the Strategic innovation program Medtech4Health, jointly supported by VINNOVA, Formas and the Swedish Energy Agency.

Sections

Disclaimer

This document does not constitute legal advice.

Contents

Context

The most important enabler for world-class computational research in biomedicine is access to massive amounts of high-quality data. OpenScience and FAIR data sharing are means to achieve this. However, despite detailed data protection guidelines and support functions, many of the implementational details are still left to the individual researcher to resolve, and effective data sharing is often hindered because of the resulting uncertainties.

The AIDA data hub has been established as a place where researchers can gather, annotate, share and enrich massive amounts of training data to support the development of artificial intelligence (AI) in medical image analysis, in a way that ensures regulatory compliance and where contributors can get increased exposure and credit for their work.

AIDA can also facilitate large scale exports for research from clinical production systems in medical imaging, and can cover costs for sharing prioritized data on the AIDA data hub. AIDA can also help researchers draft ethical review applications that include data sharing.

This section relates AIDA decisions about its data sharing activities to what AIDA has identified as the common practice of using and sharing clinical imaging data for research in Sweden and similar countries.

AIDA is a collaboration arena for AI in medical image analysis. Here, academia, industry and healthcare meet in innovation projects to translate progress in technology into patient benefit in the form of AI tools that are directly useful in day-to-day clinical care.

Common practice in the use of clinical imaging data for research in Sweden

The general data protection regulation (GDPR) regulates all processing of personal data in Europe. It however allows countries to adopt national legislation in specific well defined areas. Healthcare and research are two such areas.

Clinical imaging data is typically collected to provide care for patients. Swedish patient data law (PDL) however allows caregivers to also use it for other legally specified purposes if needed, for example for improving quality and safety of care, administration, planning, follow-up, evaluation, overseeing, and to produce healthcare statistics.

GDPR also allows this data to be used for research, for example at universities or companies. The ethical/legal framework surrounding these activities is complex, however the take-home message is that most research is allowed, if only it can be properly motivated. (And the obvious corollary: if it cannot, then it is not).

In brief, the common practice is that caregivers disclose data to research institutions for specific activities described in approved ethical review applications, to be carried out under appropriate technical and organizational protective measures and supervised by a named competent researcher. The research institution is then data controller and copyright holder for the disclosed data, and is responsible for ensuring that data is processed and shared only as described in the approved ethical review application, with data processing agreements, pseudonymization, anonymization and licensing as tools, and (starting 2020-01-01) with an obligation to store relevant data for 10 years after last use for purposes of research validation.

More on appropriate protective measures and ethics in the sections below.

Appendix 1 discusses the legal context around this common practice in further detail, with reference to original sources.

Appropriate technical and organizational protective measures

As we can see in Appendix 1 and its section on Protective measures, despite detailed data protection guidelines and support functions, many of the implementational details regarding how to properly protect personal data are still left to the individual researcher to resolve. This is to a large part due to the nature of research itself, whose mission it is to investigate the unclear and to learn the unknown, which makes it hard for anyone to usefully advise researchers on how to best protect the data that they themselves are the most intimately acquainted with and our foremost experts on.

Researchers can to some degree find support from data protection officers, institutional policies, and in policy documents such as this, and perhaps more in the approval of ethical review applications (that they themselves likely wrote!) but detailed questions like "can this data be considered anonymous?", and "are these protective measures appropriate?" must by necessity be answered finally by the researchers themselves.

There are however two encouraging consequences arising from this: while the ethical/legal framework surrounding these activities is complex, most research is actually allowed if only it can be properly motivated. And the obvious corollary: if it cannot, then it is not.

More on this in the Ethics and protective measures in research on clinical data section below.

General notes

GDPR lists principles for protection, including purpose limitation (only for specified legal purposes), data minimization (only use relevant and necessary data), storage minimization (do not keep longer than necessary), and integrity and confidentiality (technical and organizational measures). It also suggests using encryption and pseudonymization when suitable. Also, GDPR requires that protective measures must be appropriate given state-of-the-art, and be followed up to ensure that this remains true over time.

It is a good practice to review each processing step in the data flow all the way from source through research to archival for research validation, and look for opportunities to apply these principles and protective measures, taking care not to choose protective measures that hinder research activities that foreseeably could add significant value to individuals and society, such as data sharing to enable further research.

For example: Must identifying data be sent, or is it possible to pseudonymize or even anonymize at the source? Can this data be anonymized for publications? Can this data be shared for further research? If it is pseudonymized? If it is anonymized? Do I really need all these different types of data to answer the research questions in this project, or could I do it with less?

But also: Are there more research questions I could potentially answer if I could get access to more types of data in the same extraction process, for example for ruling out that certain factors contribute to disease, and could that add significant value to individuals and society without significant increase in risk and effort?

And especially: At what points would it be possible for me to introduce protective measures that could significantly reduce risk without significantly hindering useful research?

Informational protective measures

It is very easy to protect data that no longer exists. These measures concern identifying points where data or parts of data can be pseudonymized, anonymized or even deleted.

Pseudonymization entails processing data so that it no longer can be attributed to a natural person, without the use of additional information (eg: a key) which is to be kept secret and separate. Anonymization entails deleting that key. It is nontrivial to assess the quality of pseudonymization/anonymization.

The AIDA definition of anonymous data is described in AIDA GDPR policy 1.0. Appendix 3 describes what AIDA considers correct anonymization in medical imaging using examples from the AIDA dataset register along with rationale to support researchers trying to choose appropriate anonymization methods for their own datasets.

Even in cases where an image is inarguably anonymous, it may still be potentially sensitive information and identifiable through its associated data, like diagnosis, lab analysis data, time stamps, patient age and sex and so on. This problem is not unique to the medical field, and a lot of work has already been done to clarify how such data can be anonymized effectively. The Finnish social science data archive have produced a very good guide on anonymization and personal data in research, including anonymization techniques (please search in-page).

Techniques (please see the FSD guide on anonymization techniques for details):

1. **Removing variables, values and units of observation**

2. **Recoding variable values**
   Specify region rather exact location, represent age not with birth date but as "number of years" or give as ranges appropriately large given the data, eg 1, 2-3, 4-6, 7-10, or 11+ years, etc...
3. **Editing responses in open-ended variables**
   Remove specific names, addresses etc in provided free-text information. Replace with [higher education], [sports], [identifier removed], etc...
4. **K-anonymity and L-diversity**
   Choose anonymization techniques that ensure that at least K (preferably at least 3, or 5-10) individuals are represented for each possible set of indirectly identifying attributes, and that there are at least L (preferably at least 2) values for each sensitive attribute. This ensures that specific sensitive characteristic cannot be attributed to an individual, or a whole group of similar individuals.
5. **Noise addition**
   Add small additive or proportional noise to numbers, like age, starting time of treatment, inter-examination time intervals, etc...
6. **Permutation**
   Scramble values of attributes among individuals. This preserves statistical characteristics within each variable on its own, but it no longer becomes possible to study correlations between variables.

A similar guide aimed at the general public has been produced by the UK information commissioner's office: ICO anonymization code of practice. Implications for UK administrative data research have also been published.

Examples: Can I extract data from a random subset of individuals rather than the complete population? Can I pseudonymize after linking data from different sources? Can I delete the identifying raw data? Can I adequately protect any encryption keys used for the raw data? Can I delete them?

Organizational protective measures

This entails demonstrating limiting access to only the authorized people, limiting the number of authorized people, having suitable information security policies, ensuring appropriate competencies and compliance, having necessary agreements in place, not using third party services that cannot give sufficient legal and technical guarantees for safeguarding the data, etc.

Examples: Can the pseudonymization key be kept in a safe, only accessible to senior project management? Can we implement project separation, so that researchers who are engaged in several projects cannot mix data from different projects by mistake? Can we use only systems operated by public authorities such as a university, which are bound by public access to information and secrecy law (OSL)? Do we have access to any systems that have already been deemed secure enough for this processing? If we must use a cloud provider or other third party for processing, can we choose a national one? Who procured it and deemed it safe for use with this kind of data, and on what basis? Does the data processing agreement cover all necessary processing? Can we guarantee that all who access data has

affirmed their identity at some point, for example by showing their passport at employment or account generation? Are they required by law or legally binding contract to not disclose data ("tystnadsplikt")?

Technical protective measures

To be effective, modern research makes frequent and powerful use of staggeringly complex technical systems (such as computers and networks) which can be made to function in weird and wondrous ways. The set of technical protective measures possible to employ is therefore similarly diverse and complex, as this concerns security in depth at many abstraction levels, and where the devil truly is in the details. Most of these measures concern limiting attack surfaces to what is necessary and possible to follow up. Commonly employed tools are encryption and access control.

Examples:

- **Transmissions** Can we encrypt our transmissions to protect against eavesdropping? Is it more safe to send the data over a secured network connection than sending it by mail?
- **Authentication and authorization** Can we send access credentials over separate connections to protect against impersonation? Can we use multi-factor authentication (username, password, project name, client certificates, verification codes by sms or TOTP, separate VPN login, etc...)? Can we use one-time or time dependent passwords to protect against replay attacks? Can we keep access logs and monitor them for irregularities?
- **Networking** Can we restrict the IP ranges that can access the systems? Can we limit the attack surface by limiting the number of services on open network ports? Can we close down everything, and then open only what we explicitly need? Can we limit Internet access to protect against malicious code? Can we disallow outgoing connections altogether? (And then open only what we explicitly need, eg for security updates)?
- **Systems** Can we reduce the attack surface by limiting the number of softwares and systems and modes of use? Can we ensure all software is from trusted sources and continuously security patched?
- **Storage** Can we encrypt data at rest to protect against theft? Can we use encrypted disks such that they cannot be read if removed from the computer? Can we use disk retention clauses (eg document and destroy broken hardware on premise rather than returning it) to reduce the risk of unlawful data recovery off-premise?
- **Encryption** Can I ensure only strong-enough encryption is used for this data? How can I demonstrate that encryption keys are adequately protected.

Ethics and protective measures in research on clinical data

As noted in the preceding section, it is a good practice to choose protective measures such that they do not hinder research activities that foreseeably could add significant value to individuals and society.

Taking data sharing in medical imaging research as an example: extraction of clinical imaging data for research requires significant effort from caregivers that could be otherwise be spent treating the sick, so there exists a solid ethical argument that once extracted we have a moral obligation to use the data as well as we can, as extensively as possible, and in as many research projects as possible, in order to give the best possible value to individuals and society in the present and future for the effort invested.

Such a stance of course requires that researchers take every care to employ excellent and appropriate protective measures to reduce any resulting risk to individuals and society. This balancing of risk against public interest and protective measures put in place is what is evaluated in an ethical review application to the Swedish ethical review authority ("etikprövningsmyndigheten", EPM). If the researcher successfully defends their choices, the application is approved and provides a framework for what research activities are legally and ethically allowed.

There is however a danger of needlessly limiting the research activities by overspecifying them in the ethical review application.

Taking statistical processing as an example: the exact equations used are perhaps less important than the fact that "advanced statistical methods" will be used, and neither of these perhaps constitute risks that need to be defended with special protective measures in an ethical review application. Rather, it may be more pertinent to state that the exact choice of statistical method is preliminary, and then go on to describe what systems will be used to process the data and who will have access to them, how the data will be sent there, how long it will be kept, and how results will be extracted.

If it turns out that potentially valuable research activities (such as data sharing to enable further research) are not explicitly covered by language in the approved ethical review application, the researcher can seek support for these activities by motivating them in a change application to the EPM, which normally takes less time and effort than a full review.

The key lesson is to identify what are risks, and what are not risks, choose protective measures to counteract the risks, specify them, and motivate the choices. Another important take-home message is that it can sometimes be more ethical to aim higher.

The AIDA definition of anonymous data is described in AIDA GDPR policy 1.0. Appendix 3 describes what AIDA considers correct anonymization in medical imaging using examples from the AIDA dataset register along with rationale to support researchers trying to choose appropriate anonymization methods for their own datasets.

Appendix 2 holds examples of language in ethical review applications to support data sharing in research.

AIDA data sharing

AIDA policies are set continuously by AIDA management, who reports to the AIDA steering group.

AIDA supports data sharing in the following ways:

1. AIDA data hub sharing, which is the primary mode of data sharing in AIDA.
2. Large-scale data exports from clinical production systems.
3. Selective sharing only with specified research groups.
4. Support to ethical review applications that include data sharing aspects.
5. Facilitating data sharing outside of AIDA.

The primary tool for data sharing in AIDA is the AIDA PACS, however AIDA also supports self-service file sharing among AIDA members.

Overview

This section describes the process from research idea all the way to sharing research data with the greater research community for better impact and visibility, and highlights the steps where AIDA can help. AIDA can offer some support to non-members in steps 2-5 below (blue), but membership is required for major support and from step 6 on (green).

1. **Become part of AIDA**
   Several ways exist. Sharing prioritized data on the data hub is one way.

2. **Ethical review application**
   Apply to EPM for ethical review. Include data sharing aspects. AIDA can help you.
3. **Prioritized data?**
   AIDA can cover costs for data hub sharing of prioritized data.
4. **Get research data**
   Use your ethical approval to get data from caregivers and other data sources. Before disclosing data, these will assess if disclosure will lead to harm ("menprövning"). Pseudonymize and protect your data. AIDA offers support, tools and data.
5. **Anonymize data?**
   Presently, AIDA only accepts anonymous data. AIDA offers examples, support and tools.

6. **Upload data**
   Contact an AIDA system administrator to upload data to the AIDA platform. You can use the AIDA PACS to visually inspect your data. Only you can see your data.
7. **Share in AIDA**
   AIDA members you share with can see your data. An AIDA system administrator can make a working copy for download or more advanced analyses. You will be notified, or asked to approve if you have explicitly requested this.
8. **Data hub sharing?**
   AIDA encourages data hub sharing. AIDA advertises shared datasets, and makes them citable in research using DOIs.
9. **Full or sample data hub sharing?**
   In the interest of OpenScience, you should make full data available on request.
10. **Select data hub sharing license**
    AIDA promotes the AIDA BY license which allows use within AIDA.
11. **Share outside of AIDA**
    AIDA will notify you of access requests, and can facilitate contacts with requesters for data sharing options outside of AIDA, and can facilitate transfers to recipients outside AIDA based on your instructions.

Getting access

AIDA is a collaboration between partners in healthcare, academia and industry. AIDA shares data primarily within AIDA (cf Personal data and legal basis below), however researchers outside of AIDA can become part of AIDA, for example by engaging in an AIDA funded activity such as a project, fellowship or clinical evaluation, or by becoming network partner. The membership fee for network partners is low, but some fee is important as a means to establish that a meaningful commitment has been made. It can be waived if the network partner contributes by other means, for example by sharing prioritized data.

Personal data and legal basis

AIDA engages in research and innovation using data mainly from medical imaging, and thus has an obligation under the general data protection regulation (GDPR) to adequately safeguard the privacy of the individuals concerned. For this reason -and to protect against for example mistakes in large scale anonymization- AIDA engages in data sharing for research only when it is ethically approved, and when a contractual agreement is in place that includes non-disclosure of data, such as AIDA partner contracts. More details including the AIDA definition of anonymous data can be found in the AIDA GDPR policy 1.0.

AIDA can presently only host anonymous data due to terms in the data processing agreement for the present AIDA platform. At the time of writing AIDA is reworking its platform to be sufficiently secure also for sensitive personal data, at which point AIDA will be able to support further modes of data sharing.

Roles

AIDA uses the following roles in relation to sharing datasets:

- **Author:** The list of persons who are credited with the work in producing the dataset, be it in defining the selection criteria, or gathering, structuring, enriching, or annotating the data, or for leading or funding the work. This is comparable to the list of authors in an academic publication.
- **Copyright holder**: The legal entity holding the copyright ("upphovsrätt") to the dataset. This is normally given by the approved ethical review application ("godkänd etikprövingsansökan") for the research project, as the research institution ("forskningshuvudman") represented by a named competent researcher ("ansvarig forskare") under whose supervision the research activities -such as this data sharing- are taking place.
- **Contact**: Contact points for information on the dataset and for access requests. This can be any person, but usually it is a good idea to include at least the first author, the copyright holder, and some contact person from AIDA for redundancy such as the data director.

Tools

AIDA PACS

The AIDA Picture Archive and Communication System (PACS) is available here:

https://aida.medtech4health.se/ids7/

The AIDA PACS is the primary tool for data storage, visualization, interaction and sharing in AIDA. AIDA members can contact an AIDA system administrator to upload data, or download data for more advanced analyses.

AIDA has chosen a PACS as its main tool for interacting with data, since a PACS is what medical imaging diagnostics professionals use as their main tool for their everyday work in the clinic, and is therefore what imaging diagnostics AI products must integrate well with in order to feasibly add value to clinical practice. Using a PACS in this way ensures that clinicians can easily remain involved throughout any AIDA supported development activity, so that requirements can be actively driven by clinical perspective, from research idea, through implementation in a CE-marked medical device, to clinical evaluation and procurement for use in diagnosis of patients. We believe that engaging across the full value chain and connecting it in every step to as unfalsifiable needs and consequences as possible, will contribute to more relevant research and innovation, and higher impact from more reproducible science, as well as better outcomes for patients.

File sharing

The AIDA OwnCloud service for self-service file sharing is available here:

https://owncloud.aida.medtech4health.se/

AIDA members can use the AIDA OwnCloud file sharing service to store and share data with other AIDA members. OwnCloud is quite intuitive to use, so AIDA members can use it for straightforward basic sharing without system administrator assistance.

Tools for large-scale data exports

AIDA supports and can share tools for large-scale data exports from clinical production systems in medical imaging. Please see large-scale data exports below.

AIDA data hub sharing

The AIDA data hub is the primary means of data sharing in AIDA. It is used for legal and ethically approved research data sharing (cf Personal data and legal basis above). AIDA covers costs for sharing prioritized data on the AIDA data hub.

A typical scenario for AIDA data hub sharing is that a research institution as copyright holder for the prioritized research data contacts an AIDA system administrator to share the data on the data hub for use within AIDA under specific licensing terms. Thereby, the data is seen as a resource that is accessible to AIDA partners for use within AIDA.

This is in line with what we perceive as common practice in the use of clinical imaging data for research in Sweden (cf above). The ethical and legal context around this common practice is discussed in Appendix 1.

Please see the following sections for details:

1. AIDA dataset register - Citability and public exposure for shared datasets.
2. Modes of access
3. Scope and priorities
4. AIDA data sharing licenses - Commonly used terms for sharing.

AIDA dataset register

The AIDA dataset register shows public information on the datasets that have been shared on the AIDA data hub, and is available here:

https://datasets.aida.medtech4health.se/

The AIDA dataset register is used to advertise shared datasets, and to make shared datasets citable in academic research papers using Digital Object Identifiers (DOI). See Visibility, citability, and digital object identifiers below for more details.

Modes of access

AIDA members can visually inspect and interact with the data that they have access to through the AIDA PACS. For use in more advanced analyses within the dataset's sharing license, a working copy can be obtained on request to an AIDA system administrator. Copyright holders will be notified of the use, but will be asked to approve the use if they have expressly requested this.

Data shared on the AIDA data hub is by default available either in full or in part on the AIDA PACS, and shall in the interest of OpenScience be made available in full on request to the copyright holder.

Scope and priorities

AIDA can cover costs for sharing, extraction and enrichment of prioritized data on the AIDA data hub.

AIDA data priorities are continuously set by the AIDA data hub clinical council, to enable development of those AI tools that will best meet the needs of current and future clinical practice. The data priorities are being continuously updated based on current data hub

composition and identified clinical needs. Please contact the AIDA data director for proposals for data extraction or to suggest further data acquisition topics.

Visibility, citability, and digital object identifiers

The AIDA dataset register is used to advertise and provide information on datasets that have been shared on the data hub, and to make them citable in academic research papers using Digital Object Identifiers (DOI). This helps promote OpenScience, and helps make shared datasets more Findable, Accessible, Interoperable and Reusable (FAIR)

AIDA acknowledges that data collection and preparation is an important part of any data science endeavor, and that recognizing it as such will stimulate further sharing, which will lead to more impactful research and improved patient benefit.

DOIs work by means of the global DOI system, such that following a DOI link for a dataset will always take you its official landing page, which holds the latest updated information on that dataset.

Example: 10.23698/aida/ctpa → https://doi.org/10.23698/aida/ctpa

AIDA has its own DOI prefix provided by DataCite through the Swedish National Data Service (SND), and has made the data hub and its datasets discoverable through the global re3data.org registry of research data repositories, here: AIDA data hub on re3data.

AIDA data sharing licenses

For data hub sharing AIDA recommends using one of the license types that are enumerated in decreasing preference here and further described below:

1. AIDA BY license.
2. AIDA license.
3. AIDA CA / AIDA BY CA license (for limited time and until first publication).

If the terms of these licenses are not suitable for a specific case of sharing, then the copyright holder is welcome to contact AIDA to agree on more suitable terms for sharing.

AIDA license

The AIDA license is adapted from the succinct, understandable and permissive ISC license commonly used in open source software, but modified here to disallow use outside of AIDA. This restriction is a safeguard against mistakes, for example in large scale anonymization, which may otherwise lead to privacy breach.

Example:

```
Copyright <Year> <Copyright holder>
```

Permission to use, copy, modify, and/or distribute this data within AIDA (Analytic Imaging Diagnostics Arena https://medtech4health.se/aida) for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

THE DATA IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS DATA INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR CHARACTERISTICS OF THIS DATA.

AIDA BY license

AIDA recommends the AIDA BY license for most cases of data hub sharing, for reasons described in the section on Visibility, citability, and digital object identifiers above.

The AIDA BY license type builds on the AIDA license (cf above) and adds attribution requirements on publications resulting from the use of the data, for example by citing the dataset and related papers and/or including an acknowledgement text.

Example:

Copyright <Year> <Copyright holder>

Permission to use, copy, modify, and/or distribute this data within AIDA (Analytic Imaging Diagnostics Arena https://medtech4health.se/aida) for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies, and that publications resulting from the use of this data <OPTIONAL: include the following acknowledgement text “...” and> cite the following publications:

<This dataset>
<Paper 1>
<Paper 2>

THE DATA IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS DATA INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR CHARACTERISTICS OF THIS DATA.

AIDA CA license

Note: Only recommended by AIDA for limited time and until first publication.

The AIDA CA license type builds on the AIDA license (cf above) and adds a requirement that publications resulting from the use of this data include (specified) dataset authors in the author list.

AIDA only promotes using this license for limited time and for datasets where data collection is actively ongoing, such that the dataset authors can be seen to make a clear contribution to the publication. The intended effect is to increase and reward early sharing, and to stimulate forming new research collaborations, which may influence aspects of data collection toward greater generalizability to support further lines of research inquiry. Once an AIDA CA licensed dataset has been used in a publication, AIDA recommends changing the license to no longer require co-authorship, but rather citations or similar, in line with good OpenScience research tradition and ethics.

Example:

> Copyright <Year> <Copyright holder>
>
> Permission to use, copy, modify, and/or distribute this data within AIDA (Analytic Imaging Diagnostics Arena https://medtech4health.se/aida) for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies, and that publications resulting from the use of this data include the authors of this dataset <optional: NN, NN, and NN> in the author list.
>
> THE DATA IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS DATA INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR CHARACTERISTICS OF THIS DATA.

AIDA BY CA license

Note: Only recommended for limited time and until first publication.

The AIDA BY CA license type is a combination of the AIDA BY and AIDA CA licensing terms.

Example:

Copyright <Year> <Copyright holder>

Permission to use, copy, modify, and/or distribute this data within AIDA (Analytic Imaging Diagnostics Arena https://medtech4health.se/aida) for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies, and that publications resulting from the use of this data include the authors of this dataset <optional: NN, NN, and NN> in the author list and <optional: include the following acknowledgement text "..." and> <optional: cite the following publications:

<This dataset>
<Paper 1>
<Paper 2>

THE DATA IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS DATA INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR CHARACTERISTICS OF THIS DATA.

Large-scale data exports

AIDA can support large-scale exports of medical imaging data for research. Researchers and caregivers are welcome to contact AIDA with any questions relating to this matter.

Such exports are often problematic for caregivers and research institutions for practical reasons since data may be large, possibly sensitive, and hard to transfer without affecting the performance of networks and equipment that may be critical for patient well-being.

Engaging with AIDA for large scale exports can be a way for caregivers to ensure that a well understood regulatory compliant procedure is used, and that critical operations will not be disrupted.

The recommended process is to transfer the data for AIDA data hub sharing if possible, to enable more research groups to ethically and legally utilize the data in further research, thereby generating more value for research and the clinic for the same effort spent in data extraction.

AIDA can also support exports for selective sharing if data hub sharing is not possible.

AIDA also offers advice and shares tools to support exports to other destinations.

Selective sharing

AIDA members can upload private data to the AIDA platform, and can choose to share data selectively with other specified research groups in AIDA at own cost. This can be done either through the AIDA PACS by request to an AIDA system administrator, or by using the AIDA self-service file sharing service.

Support to ethical review applications in research

AIDA can give support to researchers in drafting ethical review applications that cover data sharing in medical imaging diagnostics. Appendix 2 holds some suggestions on language that can be used. Researchers are welcome to contact AIDA with any questions relating to this matter.

An approved ethical review application defines the framework for what is allowed activities in most research on humans and personal data. Appendix 1 discusses the legal/ethical context around this. Applications are submitted in Swedish to the Swedish Ethical Review Authority ("etikprövningsmyndigheten", EPM). The mechanisms for submission have varied. Some of the sections are relevant for data sharing.

Facilitating data sharing outside of AIDA

AIDA can support researchers who as copyright holders wish to share their research data outside of AIDA, and can facilitate contacts between data requesters and copyright holders for data sharing options outside of AIDA.

AIDA can facilitate data transfers from AIDA systems to recipients based on instructions from the copyright holder. AIDA can also advise and share tools for data sharing.

Appendix 1
Discussion on law, ethics and the common practice in the use of clinical imaging data for research in Sweden

The legal and ethical framework surrounding the use of personal data in research is complex. Much depends on the specifics of the particular details and the context surrounding each individual case, suggesting that it would be more appropriate to give recommendation on a case-by-case basis than for a hypothetical general case. Therefore, this is not an in-depth analysis of the framework, but a limited and superficial one. In the interest of readability, many potentially important considerations such as secrecy, law enforcement and national security have been left out of this discussion, although they may be very important to whatever case you as the reader may currently have in mind. Likewise, depending on circumstances, there may exist venues for legal and ethical research that are not covered in full here, such as consented data collection, clinical studies approved by the medical product agency, medical examinations as part of the ethically approved research activities, or through biobanking or register based research.

This section is a discussion on the ethical/legal framework around the common practice in the use of clinical imaging data for research in Sweden, with reference to official sources which may be useful as starting points for someone looking to understand the framework, that they may become better equipped to evaluate their own engagements.

This discussion is based on European law, and Swedish law in places where European law leaves room for national legislation. Researchers in countries similar to Sweden may likely be able to use this document as a point of departure for their own discussions, and may have similar legislation that can be used in places where this document references Swedish law.

Note: This document does not constitute legal advice.

In Sweden the use of clinical data in research is regulated by:

- Dataskyddsförordningen (GDPR): European General Data Protection Regulation.
- Patientdatalagen (PDL): Patient data law.
- Tryckfrihetsförordningen (TF): Freedom of the press act.
- Offentlighets- och sekretesslagen (OSL): Public access to information and secrecy act.
- Etikprövningslagen (EPL): Ethical review act.
- Upphovsrättslagen (URL): Copyright law.
- Lag om ansvar för god forskningssed och prövning av oredlighet i forskning: Good research practice and review of research misconduct act.

- Institutional policies, describing how institutions shall act and operate in order to uphold for example ethics, privacy and data protection obligations.

Explicitly, the Swedish biobank law does not regulate the use of clinical data in research, but rather concerns the use and keeping of physical human biological materials, and defines rules that are much more strict than the above laws that regulate data processing. The possible reason is that a tissue sample can potentially be analysed in a myriad ways that could give rise to a potentially huge amount of sensitive personal data, whose consequences of disclosure could be very hard to assess.

The General Data Protection Regulation (GDPR) is the law that regulates all processing of personal data in all of Europe, which according to GDPR Article 4 §2 includes a very broad range of activities such as collection, organisation, storage, alteration, retrieval, use, disclosure, alignment, restriction, or even erasure and destruction of data. GDPR Article 6 §1 requires all processing of personal information to have a legal basis and be for a specified purpose, and GDPR Article 5 §1 b states that further processing for research is not to be seen as incompatible with any initially defined purposes. However, most data from medical imaging concerns information on the health of a natural person, which is a special category of information given special protection by GDPR Article 9 §1, which states that processing these categories of information is forbidden, except in a defined set of circumstances.

GDPR Article 9 §3 allows caregivers to process patient data according to national law, which in the case of Sweden is the Swedish patient data law (PDL). PDL Chapter 3 §1 requires caregivers to maintain patient health records ("patientjournal"), and PDL Chapter 2 §6 as well as GDPR Article 4 §7 identifies the caregiver as data controller ("personuppgiftsansvarig") to these records. These records may be electronic, and may be distributed over several systems (eg PACS, LIS, etc). PDL Chapter 3 §2 identifies these records as a source of information for research.

Swedish freedom of the press act ("tryckfrihetsförordningen") TF Chapter 2 §4 states that documents at a public authority ("myndighet", eg a caregiver) are public, and TF Chapter 2 §1 states that all shall have the right to access them, unless according to TF Chapter 2 §2 access must be restricted for example on account of individual privacy.

GDPR Article 9 §2 g allows research processing of such information according to national law, if it is pursuing a substantial public interest, is proportional to the aim, and has adequate protections in place according to GDPR Article 32 and GDPR Article 89. In Sweden this is established according to the Swedish ethical review act (eg "etikprövningslagen") EPL §6 in an application for review to the Swedish Ethical Review Authority ("etikprövningsmyndigheten", EPM). An approved ethical review application then gives the framework for the research activity in terms of what data processing is allowed, and what safeguards are to be employed (cf Protective measures below). This framework can be modified by submitting an amendment application ("ändringsansökan") to EPM, if needed for example to get support for further kinds of processing.

The ethical review application specifies the legal entity under whose control the research activities will be carried out ("research institution", cf EPL §2 "forskningshuvudman"). In Sweden this is nearly always an organization, such as a university or a company. EPL §11 allows research only if it is carried out by or under supervision by a researcher ("ansvarig forskare") who has the necessary scientific competencies, which includes the awareness and ability to ensure that the activities are carried out as described in the ethical review application and according to institutional policies at the research institution.

According to Swedish public access to information and secrecy act ("offentlighets- och sekretesslag") OSL Chapter 6 §4 caregivers shall on request disclose the documents ("lämna ut handlingarna") to the research institution, unless for example according to OSL Chapter 21 §7 it can be assumed that the disclosed information will be processed in breach of EPL. This means that the caregiver must not disclose information outside of what is described in the approved ethical review application, nor if it can be assumed that the recipient will abuse the disclosed information for example by processing it by means or for purposes other than what is described in the approved ethical review application. Also, OSL Chapter 25 §1 and OSL Chapter 21 §1 require the caregiver to assess whether disclosure could harm an individual ("menprövning"). This should take into account that research institutions that are public authorities (such as universities) are also required by OSL to assess harm before any subsequent disclosure, whereas private companies are not. Such assessments are typically carried out with support from internal institutional policies and guidelines for information classification and management.

If the caregiver discloses personal information for use in research activities under a research institution's control as described in the approved ethical review application, then the research institution becomes data controller to the disclosed information, and is from then on responsible for the legal and ethical processing and safeguarding of the disclosed information.

The research institution can according to GDPR Article 28 §3 also make use of third party services for processing personal data (cf Data processing services, and the cloud below).

The research institution also has the copyright ("upphovsrätt") to the disclosed information according to Swedish copyright law ("upphovsrättslagen") URL §49, because it has "produced a catalog, table or similar into which large quantities of information have been put together", since the data extraction is carried out according to parameters specified in the approved ethical review application. If the approved ethical review application allows, the research institution can also grant use of the information to others, and can use licensing terms ("avtalslicens") to further limit what use is to be considered allowed according to URL Chapter 3a.

Starting 2020-01-01, the research institution will then be required to store the information for 10 years after last use, in order to enable research validation according to §8 of Swedish Good research practice and review of research misconduct act ("lag om ansvar för god forskningssed och prövning av oredlighet i forskning").

Data processing services, and the cloud

The research institution can make use of third party services also for processing personal data according to GDPR Article 28 §3 if a data processing agreement ("personuppgiftsbiträdesavtal") is in place which describes what processing is allowed, however such an agreement may according to GDPR Article 28 §1 only be made with a data processor ("personuppgiftsbiträde") that can provide sufficient guarantees that appropriate technical and organisational measures are put in place (cf Protective measures below). The data processor may also according to GDPR Article 28 §3 a only process information according to documented instruction from the data controller. According to GDPR Article 82 §2 the data controller is liable for any damages caused by this processing, however the data processor is liable for damage caused by any processing it has carried out outside of these instructions.

In regards to sufficient guarantees, it may be easier for a research institution in the EU to obtain such guarantees from data processors that are based in the EU (or even in the same country) and as such themselves are regulated by GDPR, keeping in mind that data processors based outside the EU may not be legally allowed to fulfill any guarantees given, depending on national legislation in the country where they are based, and that a research institution in the EU may not have the same ability to successfully pursue legal action in countries outside the EU in case of failure to uphold these guarantees.

Protective measures

Principles for protection shall according to GDPR Article 5 §1 include purpose limitation (only for specified legal purposes), data minimization (only use relevant and necessary data), storage minimization (do not keep longer than necessary), and integrity and confidentiality (technical and organizational measures).

GDPR Article 32 §1 requires data controllers to implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons. The preceding sentence contains many qualifiers that are hard to quantify objectively, like "appropriate", "taking into account", "state of the art", "cost", "nature", "risk", "likelihood", and "severity". Advise on their interpretation is provided by national supervisory authorities ("tillsynsmyndighet", Datainspektionen in Sweden) according to GDPR Article 57 §1c, and by data protection officers ("dataskyddsombud") at large organizations such as public authorities ("myndighet") according to GDPR Article 39 §1, for example through institutional policies or on request.

GDPR Article 32 §1a suggests encryption and/or pseudonymization when suitable. Pseudonymization, according to GDPR Article 4 §5 entails processing of personal data in such a manner that it can no longer be attributed to a specific person, without the use of additional information (eg "a key") which is to be kept protected and separate. If this

additional information is deleted, so that it is no longer possible to attribute the data to a specific person even indirectly, then the data is anonymous and ceases to be personal data regulated by GDPR, and thus may no longer require as high levels of protection. However, GDPR Article 32 §1d requires regular follow-up of the effectiveness of the protective measures, such as anonymization carried out in the past, since as the "state of the art" progresses and more information becomes available, it may over time become possible or even trivial to again attribute the data to a specific person, making the data anonymous no longer. Depending on how the supposedly anonymous data has been disseminated, it may then require significant effort across many organizations to again restore it to an appropriate level of protection.

As can be seen in this and the preceding section, despite detailed data protection guidelines and support functions, many of the implementational details are still left to the individual researcher to resolve. This is in no insignificant part due to the nature of research itself, whose mission it is to investigate the unclear and to learn the unknown, which makes it hard for anyone to usefully advise the researcher on how best to process the data that they themselves are the most intimately acquainted with and our foremost experts on.

Researchers can to some degree find support from data protection officers and in policy documents such as this, and in the approval of ethical review applications that they themselves likely wrote, but detailed questions like "can this data be considered anonymous?", and "are these protective measures appropriate?" must by necessity be answered finally by the researchers themselves.

There are however two encouraging consequences arising from this: while the ethical/legal framework surrounding these activities is complex, most research is actually allowed if only it can be properly motivated. And the obvious corollary: if it cannot, then it is not.

Appendix 2
Examples of language in ethical review applications to support data sharing in research

AIDA can give support to researchers in drafting ethical review applications that cover data sharing in medical imaging diagnostics. This section holds suggestions on language that can be used.

These phrases are all taken out of their respective contexts where they describe circumstances and appropriate protective measures in their respective research activities. They are not intended to be used as-is, but likely need to be adapted to specific circumstances in new ethical review applications.

Researchers are welcome to contact AIDA with any questions relating to this matter.

An approved ethical review application defines the framework for what is allowed activities in most research on humans and personal data. Some of the sections are relevant for data sharing. Appendix 1 discusses the legal/ethical context around this. Applications are submitted to the Swedish Ethical Review Authority ("etikprövningsmyndigheten", EPM). The mechanisms for submission have varied. At the time of writing an online processing system is being replaced, so in the interim forms for manual processing.

Applications to EPM are processed in Swedish, but for the benefit of non-Swedish speaking readers, a translation to English is available at the end of this appendix.

Examples

Relevanta avdelningar

- 6. Datainsamling
    - 6.4 Hur kommer insamlad data att hanteras och förvaras?
- 7. Etiska överväganden
    - 7.1 Vilka risker kan ett deltagande medföra för de forskningspersoner som ingår i forskningsprojektet?
    - 7.2 Vilken nytta kan ett deltagande medföra för de forskningspersoner som ingår i forskningsprojektet?
    - 7.4 Beskriv hur projektet har utformats för att minimera riskerna för forskningspersonerna
- 9. Information och samtycke
    - 9.1 Kommer forskningspersonerna att informeras om forskningsprojektet och tillfrågas om de vill vara med eller inte?
        - 9.1.2 Motivera varför forskningspersonerna inte ska informeras och tillfrågas

- 10. Registeruppgifter
    - 10.1 Kommer projektet att begära ut uppgifter från ett befintligt register?
        - 10.1.2 Vilka uppgifter kommer att begäras ut och varför?
- 13. Redovisning av resultat
    - 13.1 Hur garanteras tillgång till data för forskningshuvudmannen och medverkande forskare?
    - 13.2 Vem eller vilka ansvarar för databearbetning och skriftlig redovisning av resultaten?
    - 13.3 Hur och när planeras resultaten att offentliggöras?
    - 13.4 På vilket sätt garanteras forskningspersonernas rätt till integritet när materialet offentliggörs?

Datamängder

"Den statistiska styrkan avgör det förväntade antalet signifikanta resultat och är relaterad till sannolikheten att ett visst signifikant resultat reflekterar en sann effekt. Inom OMRÅDE vet vi nu att om vi ökar antalet patienter med N nya fall, så leder det till U-V nya signifikanta fynd. Om vi dubblerar storleken på TIDIGARE_STUDIE, så bör vi då identifiera X-Y nya MÖJLIGA_ORSAKER till SJUKDOM, vilket skulle innebära ett viktigt bidrag med ny kunskap om den underliggande etiologin bakom SJUKDOM."

"Antalet bilder som kommer att extraheras relateras till den power och relevans som avses per studie och vi önskar kunna göra ett bra randomiserat urval varför antalet tar höjd för STORT_ANTAL bilder baserat på den digitala 5 årsproduktion som finns i det digitala arkivet (totalt cirka VÄSENTLIGT_STÖRRE_ANTAL bilder)."

"Många sätter idag sin tilltro till att olika stödsystem baserade på artificiell intelligens (AI) ska kunna erbjuda OMRÅDE en hjälp att både effektivisera och förbättra kvalitéten i YRKESVERKSAMMAS arbete. En utmaning i sammanhanget är att den träningsdata som behövs för att utveckla AI-system för dessa syften är kraftigt begränsad, om ens tillgänglig. Detta projekt syftar till att utveckla och publicera en öppen bilddatabas innehållande detaljerade annoteringar för träning av AI-system till stöd för den bilddiagnostik som genomförs av YRKESVERKSAMMA i samband med diagnos och uppföljning av patienter inom OMRÅDE."

Riskminimering

Sekretess

"Alla nyanställda på HUVUDMAN skriver på avtal om sekretess och tystnadsplikt."

"Personal som kommer i kontakt med data ger skriftlig försäkran att inte i något sammanhang bryta sekretessbestämmelser."

Ansvar och kompetens

"NAMNGIVEN_FORSKARE, ROLL vid HUVUDMAN, har ansvaret för forskningspersonernas säkerhet, vilket i detta projekt innebär att resurser finns för att hantera projektet enligt plan med avseende på anonymisering och säkerhet etc. Se intyg i BILAGA."

"Forskningen genomförs under ansvar av prefekten vid INSTITUTION, HUVUDMAN. Vid INSTITUTION finns nödvändiga resurser och expertis för att säkerställa forskningspersonernas säkerhet och integritet (datasekretess) samt forskningsprojektets genomförande, se BILAGA"

"Deltagarnas integritet skyddas utifrån ett strukturerat informationssäkerhetsarbete enligt ISO 27000. Detta innebär att arbetet styrs av policies, riktlinjer samt ett medvetande hos personalen."

Åtkomstbegränsning

"Inga obehöriga äger tillträde till lokalerna (passerkort och kod) där data och material lagras. För HUVUDMANs lokaler gäller särskild behörighet."

"Ingen data lagras lokalt. All data lagras centralt i en serverhall vid HUVUDMAN. Serverhallen skyddas enligt gängse normer för skalskydd. Alla uppkopplingar till den centrala serverhallen sker via krypterade kommunikationsförbindelser."

Begränsning till mindre känsliga datatyper

"Data som beskriver YRKESVERKSAMMAS arbetssätt kommer att samlas in. Dessa kommer att vara tillgängliga bara för forskare i projektet och bara publiceras i anonym form."

Begränsning till enbart data

"Inga avvikelser från klinisk rutin eller ordinarie behandling kommer att ske inom projektet."

Angående tidigare erfarenheter av den använda behandlingen: "Ej relevant, eftersom projektet inte direkt kommer leda till någon ändring eller något tillägg till patienternas vård."

Deltagande i flera projekt

"Eftersom vår studie är en observationell studie och inte en interventionsstudie, så löper inte forskningspersonerna större risker om denna observationsstudie kombineras med en annan forskningstudie."

"Flera studier kan utnyttja data från projektet men forskningspersonerna påverkas inte av detta eftersom dessa studier inte medför förnyad kontakt med deltagarna."

"Ej relevant, forskningspersonerna löper ingen risk genom att vara del av detta projekt, varför kombinationseffekter med andra projekt inte kommer uppstå."

Teknik

"Krypteringsnycklar skickas separat och med oberoende överföringsmetod."

Länkad data från flera källor

"Data kommer skickas med personnummer från DATAKÄLLOR och länkas, varefter data kommer pseudonymiseras."

"Personnummer måste behållas under datainsamlingsdelen av studien och därefter kommer bilderna och hela databasen att pseudonymiseras."

Pseudonym data

"Alla data kommer att pseudonymiseras och ges slumpmässiga kodnummer, som inte kan kopplas till personernas identitet. Kodnyckel som kopplar samman kod till personers identitet kommer endast att finnas vid HUVUDMAN."

"Lagringen av data sker så att enskilda individer ej är identifierbara annat än av forskargruppen."

"Skriftliga dokument förvaras låsta och med kodnummer."

"Databaserad information kopplas till kodnummer och kan inte avslöja personernas identitet."

"Resultaten sammanställs i tabeller där inga enskilda individer kan identifieras."

"Pseudonyma uppgifter SPECIFIKATION kommer lagras i DATABAS eller liknande inom EU/EES och i enlighet med GDPR. Data i denna databas kommer att kunna användas och analyseras av forskargrupper i Sverige och utomlands. För att få tillgång till data måste forskare ansöka om ett uttag och beskriva hur data ska användas (vilket måste vara i linje med denna ansökan)."

"Säkerheten garanteras genom att deltagaren loggar in med e-legitimation (BankID) samt att säkra datorprotokoll (https) används."

"Avidentifierade data kommer att analyseras av ansvarig forskare samt medverkande forskare."

"Pseudonymiseringsnyckeln kommer krypteras och förvaras hos HUVUDMAN, åtkomligt enbart för senior ledning i forskningsprojektet. Tillhörande krypteringsnyckel kommer förvaras inlåst i kassaskåp hos HUVUDMAN. Data i pseudonym form kommer att behandlas och delas på system avsedda för behandling av känsliga personuppgifter i enlighet med GDPR, och kommer sparas för att möjliggöra forskningsvalidering. Pseudonym data som ej längre behövs kommer att anonymiseras eller gallras. Data i anonym form kommer att publiceras och delas för att möjliggöra vidare beforskning."

"Data i pseudonym form kommer att behandlas och delas på lokala IT-system enligt HUVUDMANs rutiner, och på nationella system speciellt konstruerade för behandling av känsliga personuppgifter i forskningssyfte i enlighet med GDPR, såsom SNIC-SENS Bianca."

"Om andra forskare vill granska datat och reproducera resultaten kommer de erbjudas tillgång till pseudonym eller anonym data under icke-spridningsavtal, antingen på plats vid HUVUDMAN eller med hjälp av nationella system speciellt konstruerade för behandling av känslig persondata för forskning i enlighet med GDPR."

"Vid presentation av resultat kommer dessa att vara avidentifierade och om individuella resultat presenteras kommer dessa inte vara kopplade till identifierbar information. Resultaten kommer huvudsakligen redovisas på aggregerad nivå."

Anonym data

"Patienterna informeras ej om forskningsprojektet. De ansökande forskarna drar slutsatsen att detta inte behövs, eftersom personuppgifter inte hanteras."

"Samtycke inhämtas inte från patienterna. De ansökande forskarna drar slutsatsen att detta inte behövs, eftersom personuppgifter inte hanteras."

"Projektet medför inga risker för deltagarna. Patienterna påverkas inte av projektet."

"Projektet tillför ingen direkt nytta för forskningspersonerna. Patienterna påverkas inte av projektet. Yrkesverksamma deltagare får inte direkt nytta, men kan i framtiden dra nytta av resultat i form av förbättrade verktyg och arbetssätt."

"Inga etiska problem. Förhoppningsvis kan forskningen indirekt leda till förbättrad framtida vård."

Angående forskningspersonernas rätt till integritet: "Ej relevant för patienterna. Personuppgifter från patienterna kommer aldrig behandlas i forskningen eftersom första steget är full anonymisering."

"Insamlingen och avidentifieringen kommer att ske genom en halvautomatisk procedur. En behörig individ kommer sköta insamlingen inom sjukhusets nätverk. Därefter kommer avidentifiering ske automatiskt, varpå data kommer att lagras, under lösenordsskydd, och sedan göras tillgängligt för forskare i projektet. De ansökande forskarna drar slutsatsen att detta tillvägagångssätt innebär att projektets insamlade data från patienter inte utgör personuppgifter i lagens mening eller är integritetskänsliga och därmed inte berörs av etikprövning."

"Ansökan gäller huvudsakligen data som samlats in på klinik av medicinska skäl, oberoende av forskningsstudien. Projektet hanterar inte personuppgifter från patienter, dvs data kan inte

härröras tillbaka till forskningspersonerna. Således kan projektet inte påverka patienterna på något sätt, och därför inte heller påverka deras säkerhet."

"Data kommer att avidentifieras på sådant sätt att informationen inte kan återföras till den individ det härstammar från. Information såsom namn, personnummer och telefonnummer kommer att avlägsnas alternativt ersättas med genererad data. Även indirekt identifierande information, så som remiss- och undersökningsnummer kommer att avlägsnas eller ersättas."

"Data extraheras ur det kliniska IT-systemet genom en automatisk procedur. Ett automatiskt program kommer inom sjukhusets nätverk att sköta detta, inklusive all anonymisering. Data kommer sedan att lagras anonymiserade. Anonymiserade bilddata kommer sedan att berikas med medicinska metadata, framför allt i form av annoteringar från medicinska experter över olika vävnadstyper och sjukdomsuttryck."

"Anonymiserade data avses att publiceras och göras tillgänglig för andra forskare inom området för att ge möjlighet till reproduktion av studien och därmed öka validiteten av metoden, och även för att ge möjlighet till deras egen forskning inom till exempel utveckling av bildanalysapplikationer."

"Dataskydd är ej relevant, då projektet inte hanterar personuppgifter från patienter eller känsliga uppgifter om medverkande yrkesverksamma."

"Samtliga forskare har tillgång till data och kommer alla bidra till databearbetning och publikationer, under huvudansvarig forskares ansvar."

Examples translated to English

This is a non-professional translation to English of original phrases in Swedish, provided for the benefit of non-Swedish speaking readers. Please note that these phrases address evaluation criteria defined by the Swedish national ethical review authority based on circumstances in Sweden, and that criteria and circumstances may be different in other countries.

Relevant sections

- 6. Data collection
    - 6.4 How will the collected data be handled and stored?
- 7. Ethical considerations
    - 7.1 What are the risks that participation can entail for the participants in the research project?
    - 7.2 What benefits can participation bring to the participants included in the research project?
    - 7.4 Describe how the project was designed to minimize the risks to the participants.
- 9. Information and consent

    - 9.1 Will the research subjects be informed about the research project and be asked whether they want to participate or not?
    - 9.1.2 Justify why the research subjects should not be informed and asked
- 10. Register information
    - 10.1 Will the project request information from an existing register?
        - 10.1.2 What information will be requested and why?
- 13. Presentation of results
    - 13.1 How is access to data for the research leader and contributing researchers guaranteed?
    - 13.2 Who is responsible for data processing and written reporting of the results?
    - 13.3 How and when are the results planned to be published?
    - 13.4 How is the research subjects' right to privacy guaranteed when the material is published?

Amounts of data

"The statistical power determines the expected number of significant results and is related to the probability that a certain significant result reflects a true effect. In AREA we now know that if we increase the number of patients with N new cases, it will lead to U-V new significant findings. If we double the size of PREVIOUS_STUDY, then we should identify X-Y's new POSSIBLE_ CAUSES of DISEASE, which would make an important contribution with new knowledge of the underlying etiology behind DISEASE."

"The number of images that will be extracted is related to the power and relevance referred to per study and we wish to be able to make a high quality randomized selection, which is why the number BIG_NUMBER images includes a margin based on the digital 5 year production now existing in the digital archive (a total of approximately MUCH_BIGGER_NUMBER)."

"Today many place their trust in the fact that various support systems based on artificial intelligence (AI) will be able to offer AREA help both to streamline and improve the quality of the work of PROFESSIONALS. One challenge in this context is that the training data needed to develop AI systems for these purposes are greatly limited, if at all available. This project aims to develop and publish an open image database containing detailed annotations for training of AI systems to support imaging diagnostics by PROFESSIONALS in connection to diagnosis and follow-up of patients in AREA. "

Risk minimization

Confidentiality

"All new employees at RESEARCH_INSTITUTION sign agreements on non-disclosure and confidentiality."

"Personnel who come into contact with data provide written assurances that they in no context will breach confidentiality provisions."

Responsibility and competence

"NAMED_ RESEARCHER, ROLE at RESEARCH_INSTITUTION, is responsible for the safety of the research subjects, which in this project means that resources are available to manage the project according to plan with regard to anonymization and security etc. See certificate in ANNEX."

"The research is carried out under the responsibility of the head of the DEPARTMENT, RESEARCH_INSTITUTION. The DEPARTMENT has the necessary resources and expertise to ensure the safety and integrity of the research subjects (data confidentiality) and the implementation of the research project, see ANNEX"

"The integrity of the participants is protected on the basis of a structured information security work in accordance with ISO 27000. This means that the work is guided by policies, guidelines and staff awareness."

Access restrictions

"No unauthorized person has access to the premises (access card and code) where data and materials are stored. Special access rights apply to DEPARTMENT premises. "

"No data is stored locally. All data is stored centrally in a server room at RESEARCH_INSTITUTION. The server room is protected according to common standards for perimeter protection. All connections to the central server room are encrypted."

Limitation to less sensitive data types

"Data describing the working methods of PROFESSIONALS will be collected. These will only be available to researchers in the project and will only be published in anonymous form."

Limitation to only data

"No deviations from clinical routine or regular treatment will occur within the project."

Regarding past experiences of the treatment used: "Not relevant, as the project will not directly lead to any change or addition to patient care."

Participation in multiple projects

"Because our study is observational and not an intervention study, the research subjects are not at increased risk if this observational study is combined with another research study."

"Several studies can utilize data from the project, but the research subjects are not affected by this because these studies do not entail renewed contact with the participants."

"Not relevant. The research subjects are not at risk by being part of this project, so combination effects with other projects will not occur."

Technology

"Encryption keys are sent separately and with an independent transfer method."

Linked data from several sources

"Data will be sent with social security numbers from DATA_SOURCES and linked, after which data will be pseudonymized."

"Social security numbers must be retained during the data collection part of the study, after which the images and the entire database will be pseudonymized."

Pseudonymous data

"All data will be pseudonymized and given random code numbers which cannot be linked to individual identities. The key linking codes to individuals will be stored only at the RESEARCH_INSTITUTION."

"The data is stored such that individuals are not identifiable other than by the research team."

"Written documents are stored locked and with code numbers."

"Database information is linked to code numbers and cannot reveal individual identities."

"The results are compiled in tables where no individual can be identified."

"Pseudonymous data SPECIFICATION will be stored in DATABASE or similar within EU / EEA and in accordance with GDPR. Data in this database will be used and analyzed by research groups in Sweden and abroad. To access data, researchers must then apply for data to be disclosed, therein describing how data will be used (which must be in line with this application). "

"Security is guaranteed by participant login with e-identification (BankID) and using secure protocols (https)."

"Unidentified data will be analyzed by responsible researchers as well as contributing researchers."

"The pseudonymization key will be encrypted and stored at HUVUDMAN, accessible only to senior management in the research project. The associated encryption key will be stored locked in safes at HUVUDMAN. Data in pseudonymous form will be processed and shared on systems intended for processing sensitive personal data for research in compliance with GDPR, and will be stored to enable research validation. Pseudonymous data that is no longer needed will be anonymized or deleted. Data in anonymous form will be published and shared to enable further research."

"Data in pseudonymous form will be processed and shared on local IT systems according to HUVUDMAN's routines, and on national systems specially designed for processing sensitive personal data for research purposes in accordance with GDPR, such as SNIC-SENS Bianca."

"If other researchers want to review the data and reproduce the results, they will be offered access to pseudonymous or anonymous data under non-disclosure agreements, either on site at HUVUDMAN or using national systems specifically designed for processing sensitive personal data for research in accordance with GDPR."

"When presenting results, these will be unidentified, and if individual results are presented, they will not be linked to identifiable information. The results will mainly be reported at an aggregated level."

Anonymous data

"Patients are not informed about the research project. The applicant researchers conclude that this is not needed, since personal data is not processed."

"Consent is not obtained from the patients. The applicant researchers conclude that this is not needed, since personal data is not processed."

"The project does not subject participants to risk. The patients are not affected by the project."

"The project brings no direct benefit to the research subjects. Patients are not affected by the project. Professional participants will not benefit directly, but may in the future benefit from results in the form of improved tools and working methods."

"No ethical problems. Hopefully, research can indirectly lead to improved future care."

Regarding the data subjects' right to privacy: "Not relevant to patients. Personal data from patients will never be processed in research because the first step is full anonymization."

"The collection and de-identification will be done through a semi-automatic procedure. An authorized person will manage the collection within the hospital network. Thereafter, de-identification will be done automatically, whereupon data will be stored, under password protection, and then made available to researchers in the project. The applicant researchers draw the conclusion that this approach means that the data processed in the project, though originally collected from patients, does not constitute personal data in the sense of the law or is sensitive to integrity, and is thus not affected by ethical review. "

"The application mainly applies to data collected at the clinic for medical reasons, regardless of the research study. The project does not handle personal data from patients, ie data cannot be traced back to the data subjects. Thus, the project can not affect patients in any way, and therefore does not affect their security."

"Data will be de-identified in such a way that the information cannot be traced back to the individual from which it originated. Information such as name, social security number and telephone number will be removed or replaced with generated data. Indirectly identifying information, such as referral and examination numbers will also be removed or replaced. "

"Data is extracted from the clinical IT system through an automatic procedure. An automated program will handle this within the hospital network, including all anonymization. Data will then be stored anonymously. Anonymized image data will then be enriched with medical metadata, especially in the form of annotations by medical experts on various tissue types and disease manifestations. "

"Anonymized data is intended to be published and made available to other researchers in the field to provide the opportunity for reproduction of the study and thereby increase the validity of the method, and also to provide the opportunity for their own research in, for example, the development of image analysis applications."

"Data protection is not relevant, as the project does not handle personal data from patients or sensitive information about participating professionals."

"All researchers have access to data and will all contribute to data processing and publications, under the supervision of the named competent researcher."

Appendix 3
Examples of anonymization in medical imaging

This document contains examples of what AIDA considers correct anonymization along with rationale, to support researchers trying to choose appropriate anonymization methods for their own datasets. The intention is for this document to evolve in open discussions over time, with the purpose of establishing a national common view on anonymization in medical imaging, which has the potential to greatly facilitate medical imaging research in Sweden and beyond.

As we learned in the section Protective measures in Appendix 1, GDPR does not require absolute certainties -which is often difficult or impossible to establish- but appropriate measures relative to the risk. The AIDA stance on appropriate measures is described and motivated in the AIDA GDPR policy 1.0 section Definition of anonymous data. Essentially, this states that thoughtful anonymization is an appropriate protective measure for using clinical imaging data in research. This statement is not controversial, but supported by a veritable mountain of evidence in the form of images that are regularly and continually published openly in scientific journals as part of good clinical research practice.

What then is appropriate anonymization in medical imaging?

While we have concluded several times already (cf for example here) that this question must in each case be finally answered by the researchers themselves, we can provide a few examples of what AIDA considers correct anonymization, to support researchers trying to make thoughtful decisions regarding their own datasets.

Each case below has been deemed appropriately anonymous for AIDA data hub sharing, with select images published openly as examples on the AIDA dataset register.

General guidelines

While anonymization must always be assessed on a case-by-case basis, we can condense a few guidelines based on links and examples available in this section.

Associated information

Even in cases where an image is inarguably anonymous, it may still be potentially sensitive information and identifiable through its associated data, like diagnosis, lab analysis data, time stamps, patient age and sex and so on. This is discussed in the section on Informational protective measures.

These combinations of data types can be anonymous (depending on specific circumstances which must be evaluated on a case-by-case basis):

1. **Images.**
2. **Images + scanner model.**
   Even if the caregiver might be identified by the combination of scanners.
3. **Images + scanner model + time period.**
   Even if examinations were likely carried out in the interval between scanner model launch date and dataset creation date.
4. **Images + scanner model + time period + examination type.**
   Even if the image data is from an identifiable request type with a well defined examination protocol.
5. **Images + scanner model + time period + examination type + diagnosis.**
   Even if a low percentage may be found positive/negative in this type of examination.
6. **Even further characterizing data**
   But you must motivate your decision yourself!

Anonymization measures for clinical imaging data

In the clinical PACS, the pixel data is stored together with large amounts of related information as structured elements in the extensive and very expressive DICOM format. Many of these elements may contain identifying information. It can be nontrivial to know which need to be removed or modified in order to ensure proper anonymization. Different approaches are differently effective for different datasets, so for each case one should strive to select a method that is satisfactory in terms of anonymization while still not negatively impacting the ability to carry out the planned research. For large scale anonymization it is good practice to use automated tools to do most of the work, and to iteratively review the output and tweak the parameters, until no more identifying information is found in the results.

Elements to remove:

- Response: remove number, names and initials (pat and doc), HSAID.
- Images: remove scanned images, images with unknown source, or images with low intensity variance (usually these are pictures of text, which may contain identifiable information).

Elements to modify:

- Times: initial additive timeshift, and proportional timeshifts of inter-examination interval.

Computed Tomography Pulmonary Angiography (CTPA) data

Dataset: https://doi.org/10.23698/aida/ctpa
Example images:

30 clinical routine CTPA examinations, performed on a Philips Brilliance 64 CT or GE Lightspeed VCT. 14317 axial images (1 or 0,625mm) plus additional reformats. 5 of the CTPAs are positive for pulmonary embolism and have all the emboli carefully delineated by an experienced radiologist.

These CTPA image stacks were originally obtained in clinical routine examinations to provide care for patients. They were then extracted from the clinical PACS for research purposes and anonymized at the clinic using a programmable hardware module configured for this purpose and set up as a teleradiology destination onsite at the clinic. AIDA has developed software tools that work similarly.

In the clinical PACS, the pixel data is stored together with large amounts of related information as structured elements in the extensive and very expressive DICOM format. This data was anonymized by carefully deleting or modifying elements using automated tools, and reviewing the results afterwards for unexpected findings. Different methods are differently effective for different datasets, and several automated methods were evaluated before a method was found that was deemed adequate for this dataset, while still being feasible at scale and producing useful output for research.

For example, one of the methods that was deemed insufficient left very accurate timestamps for the examinations in the supposedly anonymized data, which together with other easily deduced properties from the dataset (eg: caregiver) and image data (eg: gender) could make reidentification by a lay person or motivated intruder plausible. In contrast, the chosen method allowed timeshifting the examinations by an arbitrary number of days, making the examinations chronologically consistent for each patient, but making their identities harder to deduce.

The fine tuning of the anonymization method nevertheless required multiple iterations of automated filtering and manual review to identify and mitigate all elements with potentially identifying information. This also had to be repeated for each scanner, as different instrument manufacturers tend to favor different types of elements for their recording of the same types of information. Notably, several elements were found to contain very accurate examination timestamps, and some images (such as the radiation dose report) contained burned-in patient information and thus could not be included in the export.

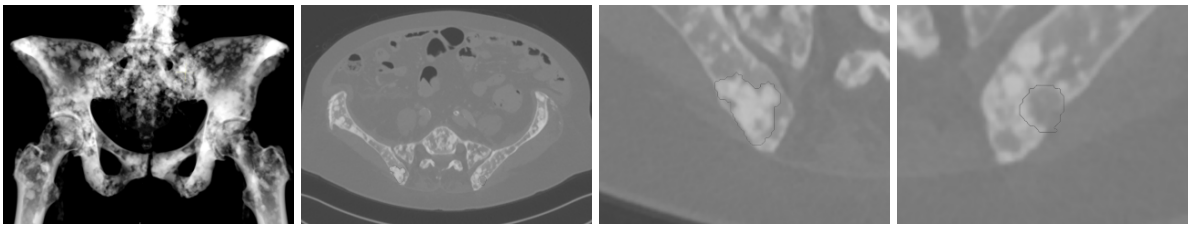
When full review of all exported DICOM data no longer revealed any plausibly identifying information for pilot cases, the configuration was deemed fit for export at scale. When exporting at scale, one examination out of every 100 were again selected for complete review of all exported DICOM data, to confirm the validity of the method, and to verify that no new types of potentially identifying information was now being added to the allowed elements (eg by clinical staff or otherwise).

It is reasonable to expect that every combination of caregiver and instrument manufacturer will have their own "house rules" for how DICOM elements are used, so it is recommended that anyone involved in anonymization carry out this type of full review of exported DICOM

data, to establish which elements are safe to disclose and which aren't, in relation to other information available on the dataset.

Skeletal data from the Visual Sweden project DROID

Dataset: https://doi.org/10.23698/aida/drske
Example images:

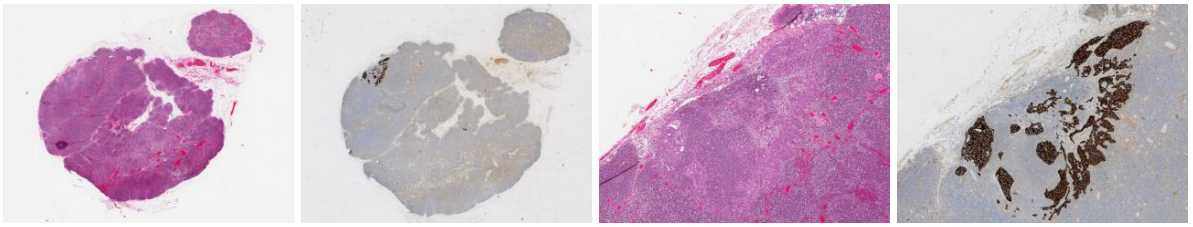


36 annotated radiology cases showing lytic and lytic/sclerotic (blastic) metastases i.e. bone regenerative and degenerative.

- Data was anonymized at the caregiver before disclosure.
- Data was collected retrospectively from routine clinical practice, representing a small part of the total number of similar examinations carried out at the caregiver.
- Data was extracted in an automatic process, where a computer program extracted the data and removed identifying information such as name, personal identification number and telephone number, or replaced it with generated information. Indirectly identifying information such as request- and examination numbers was also removed or replaced.
- Data was inspected visually by medical staff before disclosure, to ensure no identifying features were present in the data.
- This procedure was described in the (approved) research ethical review application, which also described the intent to publish and share data thusly anonymized for the purposes of research validation and as a basis for further research and AI tool development.

Axillary lymph nodes in breast cancer cases

Dataset: https://doi.org/10.23698/aida/brln
Example images:



Whole slide imaging of 396 full cases of axillary lymph nodes in breast cancer cases. Included are both sentinel node surgery and axillary resections pre, peri or post breast cancer surgery or treatment. Sentinel node cases are cut in three levels (stained with HE) and one additional slide immunohistochemically stained with CKAE1/AE3. The number of sentinel node cases with complete immunohistochemical staining is 321. The axillary

resections are cut with one cut level as default. Included are both positive and negative cases.

- Data was anonymized at the caregiver before disclosure.
- Data was collected retrospectively from routine clinical practice, representing a small part of the total number of similar examinations carried out at the caregiver.
- Case data was extracted manually from the laboratory information system. Corresponding image data was extracted from the PACS in an automatic process, where a computer program also removed identifying information such as name, personal identification number and telephone number, or replaced it with generated information. Indirectly identifying information such as request- and examination numbers was also removed or replaced. Time stamps were removed and the order of examinations was randomized.
- Data was inspected visually by medical staff before disclosure, to ensure no identifying features were present in the data.
- This procedure was described in the (approved) research ethical review application, which also described the intent to publish and share data thusly anonymized for the purposes of research validation and as a basis for further research and AI tool development.