Chapter 11
Databases and Data Retrieval

Andrea G. Rockall

Contents

11.1 Introduction

Developing a good database for imaging research is an essential task that should not be underestimated. One of the main aims of developing an accurate and user-friendly database is to support high-quality research for discovery of imaging biomarkers, biological validation of existing and novel imaging biomarkers, and model development in the machine learning domain [1]. Quality data curation is at the foundation of reliable research findings and the avoidance of false discovery. Imaging databases may range from relatively small study-specific datasets to much larger population biobanks.

DEFINITION: Imaging Biobank

Organized database of medical images and associated imaging biomarkers (radiologic and clinical) shared among multiple researchers and linked to other biorepositories.

A. G. Rockall (✉)
Department of Cancer and Surgery, Imperial College London, London, UK
e-mail: a.rockall@imperial.ac.uk

B. F. Branstetter IV (ed.), *Practical Imaging Informatics*,
https://doi.org/10.1007/978-1-0716-1756-4_11

FURTHER READING: Imaging Biobanks

ESR Position Paper: Imaging Biobanks 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4519817/

Whole-Body MR Imaging in the German National Cohort: Rationale, Design, and Technical Background 2015. https://pubmed.ncbi.nlm.nih.gov/25989618/

Regardless of the size of the database, there are some overarching principles:

1. Protection of patient privacy is fundamental: avoidance of disclosure of protected healthcare information (PHI) must be assured while retaining the scientific integrity of the data collected.
2. Common standards are important for data sharing and reuse of data, which are often expensive and time-consuming to collect. Each database should have a standardized structured model using common data elements ensuring that all data are clearly defined and categorized. This will ensure sustainability and best use of the research investment [2, 3].
3. It is important to define the storage software and the available analysis tools. Ideally, the database should be flexible enough to allow additions in the future as new tools are developed.
4. Data security, data access, and data sharing need to be managed according to information governance principles.

Scalability of high-quality image data curation, while ensuring data integrity, remains a big challenge that needs to be met in order to harness the true research potential of medical images.

11.2 Developing a Database for Imaging Research

11.2.1 Planning What Data Need to Be Collected for an Imaging Research Protocol

- Most research studies will have an ethically approved research protocol which details the research question and the outcome measures. These are usually provided in a summary table. The data points that are required for each of the outcome measures should also be detailed in the protocol. It is important to ensure that all data points required for the outcome measures are included in the database, at the outset.
- Ensure that there are clear instructions for the data collection so that the responsibilities of each party are very clear.

HYPOTHETICAL SCENARIO: Planning Data Collection

A study asks the research questions: does tumor size on CT predict progression-free survival after surgical removal of tumor?

The outcome measure of progression-free survival requires:

1. Measurement of tumor size on CT
2. Date of surgical removal
3. Definition of disease progression
   - Increase in tumor marker
   - New site of disease on CT or increase in the size of existing lesion(s) consistent with disease progression
4. Date of confirmed progression

Some examples of other questions to answer prospectively:

- Is the tumor size data point measured on original CT by reporting radiologist or will there be a central retrospective measurement?
- Is the tumor size based on maximum transverse diameter on axial image or can a sagittal or coronal reformat be used for the maximum tumor diameter?
- Is tumor size based on volume and if so how is the volume measured?
- Is the date of progression-free survival from the date of only one measure of progression (e.g., doubling of a circulating tumor maker) or are both measures required (doubling of tumor marker **and** new disease on CT)?

- Continuous review and **monitoring of data collection** is important to ensure that the data collection protocols have been understood and correctly applied. This is particularly important in multicenter data collection. It is better to discover any issues early on and to correct these than wait until the end of data collection.
- Identify ambiguity as early as possible. No protocol is perfect, and if the monitoring of data collection identifies a problem whereby there are differences in interpretation, leading to differences in data collection, then address this early on. Consider amending the data collection instructions/user manual to clarify any ambiguity and feedback to data collectors at the research sites. This may require a protocol amendment; most often this can be dealt with by clarifying data collection manual within study protocol appendices.
- Monitoring of data collection may lead to data queries which will need to be resolved throughout the course of the database development. The audit trail of data queries and data changes should be fully transparent.

DEFINITION: Audit Trail

A record of all the changes made to a database, usually with timestamps and user logs.

- It is helpful to have a database scheme which clarifies the **provenance of data** and its pathway, in order to assist with data cleaning and data queries:
  - Research site
  - Machine vendor
  - Machine version and software version
- There are major differences in imaging databases:
  - **Image repository** developed using previously acquired prospective research studies (such as in the Tumor Cancer Imaging Archive).
  - **Prospectively acquired** database such as the UK Biobank or other national imaging biobanks.
  - **Retrospective data** curated from standard of care clinical imaging. It may be curated for a disease process (e.g., breast cancer) or imaging type (e.g., mammogram or head CT). Huge unstructured datasets in imaging are currently uncommon, partly due to the large file sizes. If data mining is intended, this may take place within the clinical PACS or a large trusted research environment.

KEY CONCEPT: Prospective vs. Retrospective

Prospective data is acquired after the research question and protocol have been established, so that biases and missing data can be minimized.

Retrospective data is mined from existing sources and is more prone to bias.

DEFINITION: Big Data

Extremely large datasets that are analyzed computationally to reveal unexpected patterns, trends, and associations. Data may be structured, semi-structured, or unstructured. Data may grow exponentially with time.

- Imaging protocol and potential variations should be recorded:
  - Which images? For example, non-contrast, arterial phase or portal venous phase CT?
  - Which sequences on MRI?
  - Which time points? A study could use a single time point or multiple time points over a course of treatment or disease.
  - When can we make exceptions (e.g., if a patient has an allergy to contrast)?
- Clarify whether unprocessed imaging or processed imaging should be collected:
  - For example, in whole-body MRI, should the individual stations be used or the composed volumes?
  - Should reformats or subtracted images be included?
  - Should still or video sequences be included?
- Indicate whether the original radiology report should be included, for example, for research into natural language processing. If so, the anonymization process and linkage to the image information is an important aspect of the data collection plan.
- **Clinical metadata** collection should be detailed. Items required, system for extraction, storage, and linkage to image data need to be planned. Examples include patient age, current diagnosis, and medications.

- **Missing data** is an inevitable part of healthcare databases. This will be present in both retrospective and prospective databases. It may be due to patients declining to continue in a prospective study, missing scans during the course of a study, or data becoming corrupted. In retrospective data collection, data may be quite heterogeneous due to differences in clinical practice and patient circumstances. The plan for handling missing data should be included: it may be that missing data points can be overcome by mathematical modelling, or it may be that cases with missing data will be considered unevaluable and removed from the database, perhaps being replaced by a case that is evaluable.

11.2.2 Types of Image Databases

- Not all imaging databases are planned around a specific research protocol with a specific research question and planned outcome measures. Data warehouses are an example.
- A data warehouse aims to collect a large amount of data from one or more clinical operational system such as PACS, RIS, or EHR:
  – Data pulled from PACS or other EHR may need to undergo data cleaning and data quality check prior to being stored in the warehouse. Ideally, the data warehouse will have data integration technology and processes that harmonize and categorize data as well as applications or tools to assist researchers to use the data.

DEFINITION: Data Warehouse

A database that collects a large amount of clinical or imaging data without a defined research question or purpose. Subsets of the data can later be mined to answer newly framed questions.

- Integration of data from multiple sources into a single database may be structured, semi-structured, or unstructured.
- **Relational database**: This is a database that stores data points that are related to one another, typically in columns and rows, e.g., image data with disease category and possibly other information such as outcome may be tabulated for each subject. An example could be a database that stores thoracic CT scan findings for multiple subjects, and also documents the presence of a lung

KEY CONCEPT: Structured Data vs. Unstructured Data

Structured data has well-defined relationships and can usually be stored as rows and columns. Each element has tags or descriptors that may provide additional information. Structured data is easy to query and can be stored efficiently.

Unstructured data is not easily parsed. Examples include text, audio, and images themselves.

Semi-structured data has some of the elements of structured data (like a DICOM header or XML tags) but is not completely categorized.

cancer, the lung cancer histopathology, and the overall survival of patient from the date of diagnosis. Examples of supporting software include Microsoft **Access** and **SQL**.

- **Open-access medical image repositories**: There are many sources of open-access medical images, most of which have associated clinical metadata. These repositories provide a variety of datasets which have varying degrees of labels or annotation, providing the standard of truth. In some cases, images may be unlabelled.

FURTHER READING: Image Repositories

Many open-access repositories are listed at http://www.aylward.org/notes/open-access-medical-image-repositories.

- **Cloud database**: This is a database that runs on a cloud computing platform. The benefits include scalability, high availability, and sustainability. Data may be stored in different ways, and most cloud database providers offer a choice of database formats, often provided as SQL or other relational databases. Using one of the main providers typically offers data protection and security, encryption, backups, and updates. These should be HIPAA/GDPR compliant for use in medical databases.
- **Hybrid cloud**: Migration of a current institutional database to the cloud may require a stepwise approach, with initial migration of some aspects or applications that may benefit most from a cloud-based provision or when a new database deployment is being planned. Some legacy or traditional on-site databases may remain in use locally, thereby resulting in a hybrid system.

11.2.3 Information Governance: Approval and Anonymization

- Prior to removing data from PACS, RIS, and/or the electronic healthcare record, it is essential that all the appropriate approvals are in place, including institutional, ethical, and information governance approvals. These will vary depending on where you work:
  - In the USA, use of data will need to comply with HIPAA.
  - In the EU, use of data will need to comply with GDPR.
- Anonymization of imaging data will require the removal of patient identifiers within both the DICOM tags and on any of the images themselves (see **Chap. 8**):
  - There are several open-source tools to assist with de-identification of images, such as clinical trials processor (CTP) or DICOM Browser.
  - A challenge may be the presence of patient name burned into the actual stored images, such as in the case of many ultrasound images or in centers where stored PDF or scanned forms include patient identifiers. A quality assurance process must be in place for detecting this.
- For anonymization of radiology reports, dedicated software should be used to remove PHI. Several automated de-identification tools are available, but they are not completely reliable.

DEFINITION: PHI

Protected health information is data that is legally and ethically considered private. In the USA, HIPAA legislation defines which data elements are protected; in the EU, the GDPR regulations define this. PHI must not be revealed to anyone who is not involved in treating the patient.

DEFINITION: Pseudoanonymization

Pseudoanonymized data contains no PHI when it is viewed in the research environment. But, behind the clinical firewall, there is a lookup table that can use a unique research identifier to deanonymize the patient and acquire more data. This is also called link anonymization.

- Data may be fully anonymized or pseudoanonymized.
- You should be able to link together all the data for one subject, even though it may come from different data sources, such as multiple time points, clinical and imaging data, and outcome data.

11.2.4 Transfer of Data from PACS and EHR: Quality Assurance

- Transfer of image data is not a trivial task due to large file sizes. Transfer from the patient record, de-identification, and deposition into an image database may require considerable network bandwidth and time.
- Check that data has fully transferred, ideally using a software tool to check equivalence of pre- and post-transfer file sizes.
- Non-image DICOM objects – plan how to handle non-image DICOM data:
  - De-identification may be problematic [4].
  - How do you integrate and structure clinical data?
  - How do you quarantine data that does not fit the data structure or fails validation?
- At what point does the anonymization step take place?
  - Need to clarify the level of de-identification and DICOM tag editing.
  - Removal of some DICOM tags can strip necessary information to reproduce the study, so careful planning is required. Poorly planned anonymization may make secondary analyses impossible.
  - Ensure integrity of linkage between the subject and new subject enrollment number as well as the study/series/date following the anonymization step. There may need to be a validation step following the anonymization procedure.

  - Need to be aware of differences in DICOM conformance and application of unique identifier (UID).
  - Beware DICOM inconsistencies that may result in data being quarantined, as manual repair will be time-consuming. Document changes when tracking and validating the repair.
- Use a lookup table with a uniform format of patient enrollment numbers and unique identifiers for each imaging study/series/instance. You may know what data you wish to collect but consider how best to organize the data:
  - What data should be linked?
  - What data must be blinded from other data?
  - Organization of different modalities.
  - Organization of different dates and time series.
  - Coordination of clinical metadata with imaging data.
- **Identification of duplicates**: this can be difficult in the context of de-identified data particularly if there are different instances of de-identification of the same subject.
- Recognizing and eliminating duplicates is essential to avoid bias of a dataset due to multiple instances of a particular image which could alter analysis. This can be checked using pixel-level data.

HYPOTHETICAL SCENARIO: Duplicate entries

You have created an anonymized research database containing imaging studies. One of your subjects gets imaged at the main hospital, and then later at an outside facility. Once the data is anonymized, how will you know it's the same patient?

- **Conformance of DICOM metadata** is important to allow interoperable use of data:
  - Remember that not all institutions use DICOM headers in the same way (especially the private tags), so it may be difficult to ensure conformity.
  - Software tools for direct manipulation of DICOM errors are available, but manual editing can be burdensome for large datasets.

11.2.5 Data Processing

- **Data formatting**
  - Following anonymization of image data, it is important to store the data according to the research study or database plan. Preservation of "raw" data from PACS may be required. During the course of a study, processed data, data labels, and annotations may be added. However, the availability of the original unprocessed data is likely to be needed and should be protected.

- Retrieval from PACS usually requires de-identification of the study by changing DICOM tags, and a copy of the DICOM data is stored. However, in addition, it may be necessary to also store other file formats such as the NIfTI format. This is an Analyze-style data format to facilitate interoperable data storage and analysis, including segmentation tasks and machine learning usage.
- Conversion of DICOM to NIfTI format may be undertaken using open-source software. However, it is important to ensure that the NIfTI conversion is uniform, as there are two versions of NIfTI, the original NIfTI-1 and NIfTI-2.

KEY CONCEPT: NIfTI

The Neuroimaging Informatics Technology Initiative format is one of several alternatives to DICOM for image storage, along with Minc and Analyze. Programs meant for one file type will not work well with another, and therefore file conversion is often necessary. File types are easily recognizable by their extension (.nii for NIfTI; .dcm for DICOM). The NIfTI-2 format is an update on NIfTI-1 that allows more data to be stored. Some imaging informatics tools can convert DICOM files to NIfTI format automatically.

- In research protocol databases, the type of image processing may be known ahead of time, and the file format can be planned accordingly. However, it is essential to ensure that the database has clear version control in order to distinguish the original raw data from different versions of processed, annotated, or labelled data.

- **Data cleaning**
  - Identification of corrupt files should be automated if possible, especially for large datasets.
  - Exclude unevaluable data, such as imaging artifacts:
    Contrast failure
    Metal artifacts
    Wrong body coverage
  - Some large data collections require manual visual inspection of images prior to incorporating them into the database, e.g., the National Lung Cancer Screening trial [5, 6].
- **Data harmonization**
  - **Prospective data acquisition for planned biobank**: Ideal data collection would be harmonization of image acquisition using the same machine/technology/software version by specifically trained technicians. This may be achievable in the case of strictly controlled prospective biobank collections with highly standardized imaging protocols, for example, the UK Biobank [7, 8].
  - **Prospective data acquisition in multicenter study**: This is the next level of data in which many aspects will be harmonized by the imaging manual and protocol. However, differences in machine vendor, software versions, and

day-to-day acquisition by different technicians may result in differences in images acquired.
  - **Prospective data collection over long term**: In this case, an imaging manual and protocol may be applied but over the longer term, but there will inevitably be changes in technology and machines that will impact data harmonization.
  - **Retrospective data curation**: Most data curation is retrospective, resulting in potentially wide variations in protocol as technology advances.
  - Processing of images within a database may be required to allow similar analysis tasks to be performed. A simple example would be resampling CT volume to ensure the same slice thickness throughout a dataset. In MRI datasets, there is likely to be a need for signal intensity normalization.
  - There is a balance to be struck between very harmonized data and more heterogeneous data for image analysis. Machine learning tools generated on highly harmonized data are unlikely to generalize. However, data which is too heterogeneous may not allow successful development of machine learning tools in the initial development phase. However, some degree of retrospective data harmonization is needed in most large datasets [9, 10].

11.2.6 Software for Image Databases

- Ideally, image data should be stored in an environment that allows viewing of the images, image processing, storage of unprocessed and processed image versions, as well as any version-controlled annotations and linked clinical metadata.
- Research platforms that offer such functionality include open-source platforms, including:
  - XNAT
  - Orthanc
  - Open Health Imaging Foundation Viewer [11]
- Several commercial platforms are also available.
- Ability to run on Windows, macOS, or Linux is an added advantage, although some are designed for one or another system.
- Ability to add modules such as SQL database.
- Ability to add plug-in analysis software ensures use by a variety of researchers including radiologists and computer vision scientists.

11.2.7 Security and Safety of the Database

- Need to ensure backup (cloud or institutional server).
- Data access arrangements need to be clear and transparent (see **Chap. 19**).

- Conditions and rules must be laid out for data security (see **Chap. 8**):
  - Consider the need for data encryption.
- Legal requirements depend on country:
  - GDPR compliant in the EU
  - HIPAA compliant in the USA
- Data may be held within a Trusted Research Environment or Data Safe Haven within an institution.

DEFINITION: Safe Haven

Data Safe Havens, a.k.a. Trusted Research Environments, are highly secure data storage environments meant for researchers who need to maintain PHI for their work. There are legal standards to ensure adequate security for these databases.

11.3 Using the Database

11.3.1 Information Governance: Data Sharing and Access

- Access to the database may be open-source or controlled. This will be part of the information governance plan for data curation.
- Sites that provide data to open-source repositories will need to have appropriate institutional approval through their information governance team, with agreed parameters.
- Some open-source databases have **data access strategies** that require researchers to request access through an access subcommittee to ensure that the application to use the data are by bona fide healthcare researchers intending to undertake viable research. Clear, transparent, and fair access policies use consistent criteria, and where there is any contentious issue, access to the biobank ethics committee should be available [4, 12].
- Image databases that are limited to approved users need a system in place for allocation of username and passwords. Access needs to be user-friendly and enable appropriate data use without requiring programming skills.
- **Data-sharing agreements** and contracts may need to be in place for use of controlled data between institutions.

11.3.2 Planning Data Usage in the Context of a Study

- Database administration should be clear and transparent. Access to certain components of the database by researchers should be appropriately restricted depending on the **user role**.

- Data access may be restricted according to allocation of data:
  - Training data for model development may be widely available.
  - Testing dataset for model performance may be restricted.
  - Allocation of training and testing data should be carefully planned. For a research protocol, this allocation should be done in partnership with the study statistician.
- In some studies, **class balance** may be important in allocation to the training and test datasets. It would not be appropriate to have all image datasets with a particular finding in one or other dataset. Stratified randomization into training and test datasets is likely to ensure unbiased class balance.
- In some studies, allocation to training and testing may be based on a very simple principle such as sequential date of the data. However, it is important to consider whether there may have been a change in machine vendor, software, and imaging protocol during the period of data collection which could result in significant differences in the images over time. This may not be of concern in relatively simple datasets, such as CXR, but could be a great significance in more complex modalities such as MRI, resulting in a failure of a model to generalize.
- Image labelling or annotation tools should ideally be available within the database, but this is not always the case. Open-source tools for image segmentation may be used, such as ITK-SNAP, 3D Slicer, or ImageJ. Online platforms for image annotation and segmentation are also available, such as MD.ai.
- Linkage of the database to processing units should be available. For example, XNAT links to NVIDIA Clara, with an automated pipeline for conversion of DICOM to NIfiTI, and then conversion back to DICOM in the model output.
- Plan for data extraction, segmentations, and results.

11.3.3 Database Merging and Sustainability

- The ability for data to be merged and integrated in the future with other data repositories is an important consideration [13].
- Tools for future sustainability are in development, such as **PRISM** for The Cancer Imaging Archive (TCIA) [14].
- There are shared software algorithms and architectures with the tools required for computing, comparing, evaluating, and disseminating predictive models [15].
- Archiving of unprocessed and processed data versions in clear folder structure is essential for future use or sharing of data with future collaborators.
- The enormous time invested in developing a professionally annotated dataset should be a future resource for testing external models for the benefit of healthcare research.

FURTHER READING: PRISM

PRISM: A Platform for Imaging in Precision Medicine https://ascopubs.org/doi/full/10.1200/CCI.20.00001

PEARLS

- Plan your database from the beginning with full understanding of the intended research or clinical outcomes and purposes.
- Well-defined data structure and relationships are the key to success.
- Anonymization is difficult. You have to remove protected health information from unexpected places. You might need to go back later and deanonymize.
- Validating and curating data are key elements of database creation and management.
- Databases work best when add-on software tools can access the data without manual intervention or exporting.

References

1. Fouke SJ, Benzinger TL, Milchenko M, LaMontagne P, Shimony JS, Chicoine MR, et al. The comprehensive neuro-oncology data repository (CONDR): a research infrastructure to develop and validate imaging biomarkers. Neurosurgery. 2014;74(1):88–98.
2. Prior F, Almeida J, Kathiravelu P, Kurc T, Smith K, Fitzgerald TJ, et al. Open access image repositories: high-quality data to enable machine learning research. Clin Radiol. 2020;75(1):7–12.
3. Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, et al. The public cancer radiology imaging collections of The Cancer Imaging Archive. Sci Data. 2017;4:170124.
4. Rovere-Querini P, Tresoldi C, Conte C, Ruggeri A, Ghezzi S, De Lorenzo R, et al. Biobanking for COVID-19 research. Panminerva Med. 2020;
5. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011;38(2):915–31.
6. Clark KW, Gierada DS, Moore SM, Maffitt DR, Koppel P, Phillips SR, et al. Creation of a CT image library for the lung screening study of the national lung screening trial. J Digit Imaging. 2007;20(1):23–31.
7. Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. NeuroImage. 2018;166:400–24.
8. Bamberg F, Kauczor HU, Weckbach S, Schlett CL, Forsting M, Ladd SC, et al. Whole-body MR imaging in the German National Cohort: rationale, design, and technical background. Radiology. 2015;277(1):206–20.
9. Basu A, Warzel D, Eftekhari A, Kirby JS, Freymann J, Knable J, et al. Call for data standardization: lessons learned and recommendations in an imaging study. JCO Clin Cancer Inform. 2019;3:1–11.
10. Bauermeister S, Orton C, Thompson S, Barker RA, Bauermeister JR, Ben-Shlomo Y, et al. The Dementias Platform UK (DPUK) Data Portal. Eur J Epidemiol. 2020;35(6):601–11.
11. Ziegler E, Urban T, Brown D, Petts J, Pieper SD, Lewis R, et al. Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. JCO Clin Cancer Inform. 2020;4:336–45.
12. Conroy M, Sellors J, Effingham M, Littlejohns TJ, Boultwood C, Gillions L, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med. 2019;286(4):389–97.
13. Gedye C, Sachchithananthan M, Leonard R, Jeffree RL, Buckland ME, Ziegler DS, et al. Driving innovation through collaboration: development of clinical annotation datasets for brain cancer biobanking. Neurooncol Pract. 2020;7(1):31–7.

14. Sharma A, Tarbox L, Kurc T, Bona J, Smith K, Kathiravelu P, et al. PRISM: a platform for imaging in precision medicine. JCO Clin Cancer Inform. 2020;4:491–9.
15. Mattonen SA, Gude D, Echegaray S, Bakr S, Rubin DL, Napel S. Quantitative imaging feature pipeline: a web-based tool for utilizing, sharing, and building image-processing pipelines. J Med Imaging (Bellingham). 2020;7(4):042803.

Self-Assessment Questions

1. The data within an image is considered:
   (a) Structured data
   (b) Unstructured data
   (c) Semi-structured data

2. Protected health information does *not* include:
   (a) Name
   (b) Age
   (c) Date of birth
   (d) Location where images were obtained

3. Existing clinical data is considered:
   (a) Prospective
   (b) Retrospective
   (c) Introspective

4. A careful record of all changes made to the data in a database is called:
   (a) Audit trail
   (b) Governance
   (c) Warehousing
   (d) Validation
   (e) Curating

5. Unlike data use within the enterprise, data use between multiple institutions requires a:
   (a) Data access strategy
   (b) Business partnership
   (c) Audit trail
   (d) Data-sharing agreement

6. What is the difference between de-identification, anonymization, and pseudoanonymization?

7. What are the legal requirements in your country for privacy of protected health information? How could you help a researcher include PHI in a database if it was absolutely needed?