J Digit Imaging (2012) 25:566–569
DOI 10.1007/s10278-012-9502-6

Imaging Infrastructure for Research. Part 2. Data Management Practices

Daniel S. Marcus · Bradley J. Erickson · Tony Pan · CTSA Imaging Informatics Working Group

Published online: 19 June 2012

Abstract In part one of this series, best practices were described for acquiring and handling data at study sites and importing them into an image repository or database. Here, we present a similar treatment on data management practices for imaging-based studies.

Keywords Clinical trials · Research image database

Data Acquisition

As described in part one, a clinical research study will typically include multiple different kinds of data (e.g., labs, clinical exams, imaging) and multiple data acquisition sites. Data may also be obtained at multiple time points, often prior to, during, and after an intervention (e.g., a drug regimen, surgical procedure). The data obtained during a visit are either stored directly in an electronic form or recorded on paper and then transcribed to an electronic form. The stored values may represent individual measurements or a calculated value based on some algorithm (e.g., a calculated score on a behavioral evaluation). For imaging data, the images are generally obtained from imaging devices in electronic format, often the industry standard DICOM format. Ancillary information about the acquisition of the data is typically recorded on a paper case report form (CRF) or electronic CRF. Ancillary information would typically include such items as the data and time of the study, observations (e.g., patient moved during acquisition), time of contrast administration, and volume of contrast.

From a data management perspective, there are several key requirements for properly managing data during and immediately after acquisition:

1. Procedures and supporting software should be established for uploading/entering each type of data obtained in the study. Both efficiency and data accuracy would favor direct electronic capture without paper forms.
2. Procedures and supporting software should be established to verify that entered data comply with protocol and fall within an allowable range.
3. Procedures should be in place to ensure that data are entered within an allowable time frame.

Data Coordinating Center

Data for a multisite trial are typically stored in a centralized data coordinating center (DCC). The DCC provides several critical functions. First, it deploys and operates the infrastructure for entering and storing study data. Second, it provides standard operating procedures for how data are entered into the system. Third, it provides quality control procedures—usually a mix of manual and automated procedures. Fourth, it provides access to the

D. S. Marcus
Radiology, Washington University School of Medicine,
4525 Scott Ave, Campus Box 8225, St. Louis, MO 63110, USA
e-mail: dmarcus@wustl.edu

B. J. Erickson (✉)
Department of Radiology, Mayo Clinic, Rochester,
200 First Street SW,
Rochester, MN 55905, USA
e-mail: bje@mayo.edu

T. Pan
Center for Comprehensive Informatics, Emory University,
201 Dowman Drive,
Atlanta, GA 30322, USA
e-mail: tony.pan@emory.edu

data for study investigators. Finally, it provides training and helpdesk support during the startup and execution of the study. The DCC works closely with investigators during study design, execution, and analysis to ensure that the data management services it provides meet the requirements of the study.

Centralization of data management through a coordinating center has a number of advantages over distributed approaches, including minimizing technology and staffing requirements and simplifying analysis and distribution of the data. Some recent efforts have focused on federated data models, where data reside at each study site and are unified through layers of software [1]. While these approaches show promise, particularly for retrospective studies and ad hoc collaborations, we believe that they are at a serious disadvantage for managing controlled prospective studies. In particular, enterprise-level hardware and software must be deployed at the study sites, which is often inconvenient or impossible.

The Database

As images come off of scanning devices, the image information and much of the metadata are often combined in a single file, as prescribed by the DICOM standard. However, for data management purposes, it is important that data be stored into a more accessible form—typically a database.

The database approach required to manage imaging-based clinical studies depends largely on the scale and scope of the study. For smaller studies, a simple spreadsheet may be sufficient. However, for most multisite trials as well as larger single-site trials, an enterprise-grade database management system is necessary to properly handle the volume and complexity of data acquired in an imaging-based clinical research study. A range of plausible database architectures could be suitable to store these data, but there are several characteristics that such a system should support:

- Storage of image data, either directly in the database or via references to a file-based image archive,
- Storage of image metadata,
- Storage of derived image-based measures,
- Storage of associated non-imaging data,
- A longitudinal data model,
- Queries between data types,
- Format-independent file storage,
- Security and protocol-level authentication,
- Provenance, history, and audit trail, and
- 21 CFR Part 11-compliance (for FDA regulated trials).

Given these requirements, several candidate architectures come to mind: a single comprehensive database, multiple federated database, or one with highly distributed services with unified security. There are advantages and disadvantages to each, and the "best" solution will depend on the nature of the problem to be solved, the computing environment, and the envisioned scope and size of the desired solution. In many respects, a single database is simpler—security is unified, there is one place to look for data, and the need for high-performance networks and servers is reduced. On the other hand, a single database may not be able to scale up to the size of the problem that you need to address now, or in the future. At the other extreme, one could imagine an array of many data sources highly focused on its type of data that are connected by well- defined web services. In this case, increasing the volume as well as the scope is simpler, but developing this is complex and may be "overkill" for simple problems. Security might also be more challenging to manage. caBIG [1] is one example of a group of services that can be tied together to provide access to a wide array of data. There is also a middle ground, consisting of a few federated data sources with corresponding intermediate trade-offs: these are more scalable, but somewhat more complex to develop for.

Picture archiving and communication systems (PACS) are used in clinical environments to manage medical images. PACS are extremely efficient at providing radiologists and other clinicians with services to view images and generate diagnostic reports. Given their wide use in clinical environments, PACS would seem to be a natural solution in research as well. However, in practice, PACS are extremely limited in their support for research imaging: they lack support for research protocols, which limits organization of data and protocol-based user authentication; they are DICOM-centric and generally do not support alternative file formats widely used in research; and they do not store research metadata, derived measures, and non-imaging measures. They also have little support for integration of display methods or image manipulation tools beyond what was included at the factory. Finally, they lack a variety of required security capabilities (e.g., file and network encryption). Given these limitations, PACS are insufficient for managing imaging-based clinical research data. However, a PACS-like component that receives and stores DICOM-formatted imaging studies is likely to be an important component of an overall research imaging database solution.

Data Organization

For a relational database, it is necessary to model the relationships between different pieces of information. Failing to correctly model how investigators will want to select and retrieve the information will significantly degrade the

performance of the system. There are several likely candidates for elements of this data model, including:

- Research study (aka project, study, trial): the entity into which subjects are enrolled;
- Subject (aka patient, participant): the individual from whom data are obtained;
- Visit (aka episode of care): a single appearance of an individual at a study site;
- Examination (aka ImageSet, experiment, scan): experimental data obtained on an apparatus or instrument in a single engagement with the imaging device;
- Series (aka scan, acquisition): a group of images that are acquired in one grouping. The exact meaning will depend on the imaging device;
- Image: usually a 2D collection of pixels produced by an imaging device; and
- Channel: device-specific data source measured during a series.

Several of these terms are used ambiguously in the field and are constant sources of confusion. "Study," for example, may refer to either the encompassing research program under which a group of subjects are recruited or to a collection of images obtained from a single subject during a single entrance into an imaging device. In the broad scientific community, "study" general refers to the former, but within radiological imaging (i.e., the DICOM standard and most radiologists), it typically refers to the latter. Such ambiguous terms are best either avoided or qualified (e.g., "DICOM study") to reduce potential confusion. "Protocol" is another such term that has many different meanings in the research community. Despite these ambiguities in terminology, the conceptual units and their interrelationships are quite clear.

Database Architecture

Imaging data have a number of characteristics that guide how they are managed in a database. First, these data include two components: metadata and binary pixel data. The metadata are typically string or numeric data that represent aspects of the image's history (e.g., acquisition parameters, device serial number); this information is useful to store in a way that allows users to select interesting subsets based on these values. Relational databases are the most popular form of database technology used for these data. However, nonrelational databases provide a flexible alternative. These are often implemented using a tag value method, which alleviates the need for decisions about the relationships between data elements. These types of databases are less efficient than relational databases when the relationships are known. However, for cases where the connections are not well understood, they can provide a useful alternative. It is also possible to use hybrids—for instance to use a relational database for the information that is well structured (e.g., the DICOM information) and tags for information that is less well structured or that may be added later. This flexibility can be useful for large image archives that might be used across multiple research projects, including future ones where the research questions and data are not known.

Because the binary pixel data are seldom directly queried, this information is therefore better managed as binary objects either in the database system itself or as files on a Unix/Windows type file system, or in a document oriented system such CouchDB (http://couchdb.apache.org/) or DynamoDB (http://aws.amazon.com/dynamodb/). Because imaging data are quite large, a relational database may be paired with a well-organized file system. The two components can be tied together formally by including file path information in the database or by following common naming conventions in the database and file system. As an alternative, the data could be stored directly in the database as a binary large object though that has several disadvantages: it reduces the flexibility of how the data are stored on disk; it makes data access more complicated; and it bloats the database, likely compromising performance.

File Management

One of the main data management challenges is determining how to handle a great many files. The implementation of the DICOM standard by most of the large vendors tends to produce many files, each fairly small. Most commonly, a file is produced for each reconstructed 2D slice, even when acquired as part of a "3D" sequence. Time series data also typically are produced as a (long) series of 2D images. It is not uncommon for these studies to consist of 20,000 files. Given the proliferation of files generated during an imaging study, a robust approach for managing files is necessary. One option is to choose a file system organization that matches the organizational units described above.

The type of file system used to manage files as described above is typically a UNIX/Windows style system that can be flexibly integrated with a Network File System, a Common Internet File System, etc. These file systems have hard limits on the number of files they can address, typically in the billions, which sounds very large, but actual instance of DICOM image archives has exceeded this limit, forcing alternative solutions to be found. One option is to combine these in a computable fashion, but that limits the efficiency of accessing subsets of images. File

management platforms, such as CouchDB and Amazon Simple Storage Service (S3 http://aws.amazon.com/s3/), provide another way to address this problem, by providing an abstraction layer between the file system and the file access methods.

Data Storage

In addition to the large number of files, imaging data tend to be quite large. Mammographic images can be greater than 50 MB each, with routine radiographs more typically being about 10 MB. It is possible to compress these using either lossy or lossless compression. Lossy compression is probably not acceptable for any research; lossless compression can reduce storage needs by a factor of about 2.5:1 but does require additional computation both when storing and retrieving images. While DICOM should be the preferred form for medical images, not all data can be represented as DICOM. Annotation and Image Markup format [2], for example, provides a standard format for representing measurements and labels associated with medical images that are targeted for research.

References

1. National Cancer Institute. The Cancer Biomedical Informatics Grid. 2011; Available from: http://cabig.cancer.gov/
2. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL: The caBIG annotation and image Markup project. J Digit Imaging 23 (2):217–225, 2010