Whitepapers on Imaging Infrastructure for Research

Paper 2. Data management practices

Daniel J Marcus, Bradley J Erickson, Tony Pan, CTSA Imaging Informatics Working Group

Introduction

The use of imaging data in clinical research can provide enormous scientific benefits, but it can also entail substantial complexity. There is a general process for developing a clinical research project. Figure 1 shows a high-level description of this process. Clinical research that involves imaging follows the same general workflow. However, the inclusion of imaging into a research protocol has additional workflow considerations and complexities. By implementing standard procedures and enforcing them through software and policy, many of these complexities can be mitigated and imaging can be successfully integrated into a variety of clinical research applications. The purpose of this set of three papers is to document some of the additional workflow considerations related to imaging that is used as part of a clinical trial.

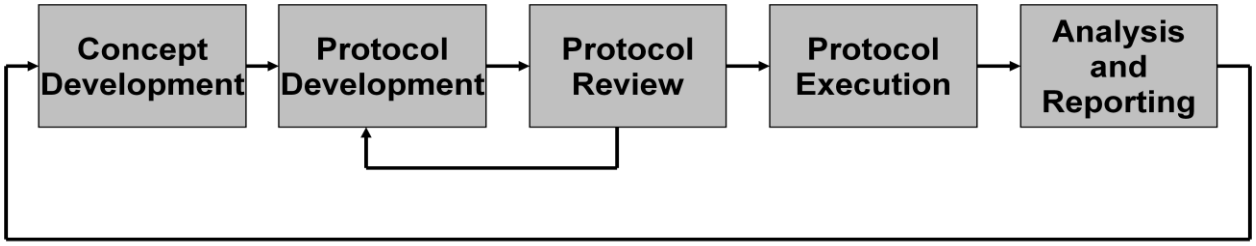

Figure 1. High level process diagram for development of a research project.

An important point that the authors wish to make is that imaging information is nearly always not simply an additional data point. The inclusion of imaging data results in many additional complexities that can lead to unintended or unrecognized biases and errors. For that reason, it is critical that imaging experts be involved in studies that rely on imaging data. That involvement is required during the conception and design of the experiment, the data collection phase, and the data analysis phase. Some of those challenges that are present when imaging data is used in research will be further defined in this paper. The proper collection of imaging data for research use demands monitoring that can be made better and more efficient than current manual methods.

How measurements are extracted from the raw images and represented is a critical step that entails its own set of challenges. Finally, there are some unique and common security issues when images are used for research.

Paper 2. Data management practices for imaging-based clinical research

In the first paper in 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, with a particular focus on integration across imaging and non-imaging data and on qualitative and quantitative assessment of images.

Data acquisition

As described in the first white paper, 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 (eCRF). 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. Software should be

deployed to support entry of each data type. 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 multi-site 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 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.

Sidebar: Why not a PACS?

Picture archiving and communication systems (PACS) are used in clinical environments to manage medical images. Given their clinical focus, PACS are extremely efficient at providing radiologists and other clinicians with services to view

images and generate diagnostic reports. However, they 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 don't support alternative file formats; nor do they 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.

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 multi-site 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.
- Stroage of image metadata
- Storage of derived image-based measures
- Storage of associatednon-imaging data.
- A longitudinal data model.

- Queries between data types
- Format-independet file storage
- Security and protocol-level authentication
- Provenance, history, and audit trail
- 21 CFR Part 11-compliance (for FDA regulated trials)

Given these requirements, several candidate architectures come to mind: one full service database, integrated/federated services, 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.

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.
- 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.

While relational databases are the most popular form of database technology used, there are alternatives that are more flexible—the nonrelational databases. 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.

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. It is less common to directly query images based on the binary pixel information. 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 or Dynamo. Because imaging data are quite large, the RDBMS may be paired with a well organized filesystem. 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 filesystem. As an alternative, the data could be stored directly in the database as a binary large object (BLOB) though that has several disadvantages: it reduces the flexibility of how the data are stored on disc; 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 2-dimensional 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 NFS, CIFS , 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 have 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 S3, provide another way to address this problem, by provide 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. Give some examples. Mammographic images can be greater than 50MB each, with routine radiographs more typically being about 10MB. 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 (AIM) is one technology that appears to be a solid method for representing measurements and labels associated with medical images that is targeted for research.

End points and other measures of interest

A wide variety of qualitative, quantitative, and semi-quantitative measures can be obtained from images. Depending on the type of study, various types of measures will be more or less appropriate. In clinical trials for FDA approval of a therapy or device, the range of allowable measures for measuring efficacy is extremely limited. These measures are often referred to as endpoints. To achieve endpoint status, the image characteristic under measurement must relate directly to clinical progression. In most cancers, the size of the tumor is measured as an endpoint, though functional imaging methods like PET sometimes use intensity measures. A somewhat broader range of measures are those that have been demonstrated in previous studies as having some power to predict outcomes. These are referred to as imaging biomarkers. One

proposed biomarker is increased cerebral blood volume on dynamic contrast-enhanced MRI images in primary brain tumors. The broadest category of useful measures are those that can be generated reliably from the images acquired in a study but have not yet been demonstrated to have predictive power. Within the imaging research community, these measures and the methods to generate them are themselves the focus of research.

While a particular measure must go through extensive validation before being raised to the level of endpoint, that does not mean that such measures are without error. Errors in generating endpoint measurements can enter at a variety of points in the imaging workflow and can be cumulative. Endpoints are generally semi-quantitative, in that they are linear or volumetric measures as completed by a trained expert who must use visual perception to complete the measures, thereby introducing significant variability. Some example endpoints are volume of tumor, volume of T2 lesion in multiple sclerosis, cerebral blood volume measurements, or an ejection fraction.

Conclusions

- There are often well-defined steps that are followed to produce results. 'Pipelines' are steps of automated image processing or extraction that can be sequenced to make computation more efficient. Workflow is the sequence of steps and decisions that humans must make to take the original input and produce the final product. Support for workflow is also critical to efficiency and reliability.
- Standardization of raw data collection and measurement representations (annotations) can improve the efficiency of data analysis as well as equivalence of results across studies.
- Seemingly small changes in steps can produce significant differences in results
- The ability to run all datasets through the workflow is often required and valuable when new insights are gained during the data collection period as well as for interim analyses.

- Efficient access to the data developed as part of a research protocol is critical the large size of images presents some unique challenges compared to most traditional research data sets. It is also important to track tools and users that created measurements, as images almost never represent the final data.
-

1. National Cancer Institute. *The Cancer Biomedical Informatics Grid*. 2011; Available from: http://cabig.cancer.gov/.

-