

# A Medical Data Collection System for Sharing Data with Outside Collaborators

Steinberg T., Wang Y., Ford J. C., Makedon F. S.

Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA  
{tilmann, wyh, jford, makedon}@cs.dartmouth.edu

## Introduction

The medical field has seen a huge increase in data collection due to new technologies and techniques, particular in imaging, where large amounts of data are generated. Medical groups are now working together with computational groups to develop methods and tools to process imaging data sets quickly and efficiently. However, due to recently introduced privacy regulations, usually only “clean” patient data can be shared; also, data owners desire to have some means of control over access to their data by collaborating groups.

We are developing a system to meet these requirements and facilitate easier data sharing. It provides a framework for developing new methods for analyzing imaging and non-imaging medical data. This system is primarily designed to work with Multiple Sclerosis (MS) data [1,2], with the long-term goal of creating exchangeable data sets from multiple sites [3].

## Prototype System Overview

Modern imaging technologies allow for a much more thorough investigation of subjects for medical purposes. As Magnetic Resonance Imaging (MRI) and similar scans become commonplace, data processing and management tools are essential in dealing with the resulting large volumes of images. At the same time, privacy regulations require additional steps before anyone outside a clinical group can have access to the data. For collaboration between clinical and outside groups, therefore, it is beneficial to design tools that ensure compliance with privacy regulations at a minimal cost to the clinical staff.

Our prototype system is designed to provide such functionality by deploying tools to automatically process incoming data and split it into parts reserved for the clinical group and parts available to collaborating groups. Furthermore, as data sets are processed, they automatically become available either for advanced processing or for inclusion in analyses. The system provides a simple overview of each component’s status, allows selection of data sets by status, and connects to processing and analysis tools to apply to such a selection. Likewise, new tools or updates to tools become available for use as they are added by the developers. Figure 1 shows a flow diagram of how a clinical and a computational group collaborate on developing tools for analysis.

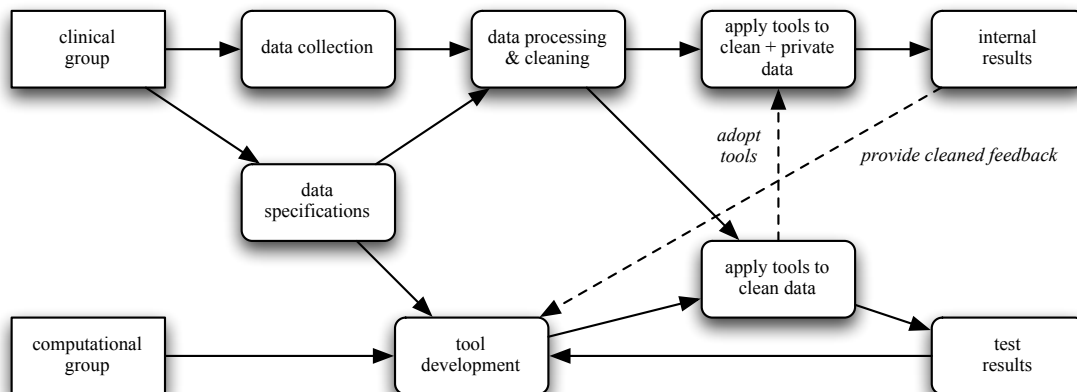


Figure 1. Collaboration between clinical and outside group (here: computational). Shown are the different steps taken to create and refine tools and results from applying these tools to existing and incoming data sets.

**System Components:** The system consists of a web-based interface, a database system for storing non-image data, a restricted file server for storing raw and processed image files, and a library of tools for data processing and analysis. Access to the system is restricted to registered users and limited to role-based functions, as defined by the data owners.

**Workflow Definition:** One central goal of the system is to allow users to have access to data and tools as they become available, and to incorporate results of processing and analysis steps automatically. To achieve this, the project leader can specify members of the project, their assigned tasks, and who should be notified

of new additions or changes. This allows for clearly delineating the roles of each participant and a tight integration of different steps, e.g. application of an updated tool to a data set.

**Data Collection:** The system expects data to come from sources such as FTP servers connected to an MRI scanner, or other regularly updated locations. Scan images consist of two parts: the raw image data and a header containing information about the scan. This header information is copied into the clinical database and then replaced by a header from which all identifying data have been either removed or replaced with internal references (e.g. subject ID number instead of name, age bracket instead of date of birth, visit sequence number instead of visit date).

**Data Processing:** Once a new image set has been acquired, processing it consists of a series of tasks that either require human interaction or decision-making, or that can be fully automated. One such task is to align brain images for the same subject so that a given location refers to the same location in the brain for all scans. This is done by manually selecting a common reference point (usually an easily discernible part), applying an alignment tool, and verifying the result. Another processing task is to strip skull and facial features from brain scans to further anonymize the data while preserving the original scan data. An alternative is to map each individual scan to a standard brain space, which can introduce distortions but ensures that the imaging data cannot be connected to the original subject.

**Using External Tools:** A large body of existing tools are based on proprietary software, such as MATLAB. Since our system is web-based, we need to create interfaces between the web application and MATLAB code. For non-interactive tools, our system creates scripts on the fly that reference the selected data, apply the appropriate tool, and read the output back into the database. Tools that require manual interaction need to be run inside a local MATLAB client next to the web interface on the user's workstation. Members of the outside group need to download a cleaned subset of the imaging data, i.e. applying tools to private data is reserved to the clinical group.

**Usage Tracking:** The system maintains a log of users' activities, such as connections, added data sets, processing steps, queries to the database, tool uploads or specifications, downloads of data sets, etc. This documents data access by the participants and allows verification of compliance with privacy regulations.

## Discussion

One main challenge of this project is to provide sufficient benefits to entice clinical users to adopt the system, without adding to their existing workload. It is our hope that the combined functionality of automating common steps that do not require manual interaction, allowing non-clinical users maximal access to data for developing and applying analysis tools, and the documentation of data usage as partial fulfillment of new regulations provide needed incentives. Fully implemented, our system would provide both clinical and outside parties with desired materials: a computational group would have access to data for testing new solutions, and the clinical group could take advantage of such new tools. However, the central advantage of the described system is that the data owner—the clinical group—has full control over which components of the collected data are made accessible to the outside group, and can track and document such outside usage. This allows the continued involvement of computational and other outside groups in developing medical analysis tools.

## Acknowledgements

This work has been made possible by funding from NSF (grants # 0083423, 0312629, and 0308229) and by the close cooperation of Dr. Andrew J. Saykin and Prof. Heather A. Wishart from the Brain Imaging Laboratory at Dartmouth-Hitchcock Medical Center.

## References

- [1] Steinberg, T., Wang, Y., Makedon, F., Shen, L., Saykin, A., Wishart, H., A Spatio-temporal Multi-modal Data Management and Analysis Environment for Tracking MS Lesions, *SSDBM 2003*, July 9-11, 2003, Cambridge, MA
- [2] Wang, Y., Steinberg, T., Makedon, F., Wishart, H., Saykin, A., Quantifying Evolving Processes in Multimodal 3D Medical Images, *Proceedings of the Sixth Annual International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2003)*, pp 101-108, Montréal, Québec, Canada, 2003.
- [3] Ye, S., Makedon, F., Steinberg, T., Shen, L., Ford, J., Wang, Y., Zhao, Y., Kapidakis, S., SCENS: a System for the Mediated Sharing of Sensitive Data, *JCDL 2003*, May 27-31, 2003, Houston, TX