Neuroinform (2010) 8:207–212
DOI 10.1007/s12021-010-9087-5

EDITORIAL

Neuroinformatics in Clinical and Translational Medicine—Novel Approaches

Randy L. Gollub · Dennis A. Turner

Published online: 6 October 2010

This special edition of Neuroinformatics was conceived as a way to highlight the new and rapidly developing field of clinical neuroinformatics. The application of neuroinformatics to the clinical domain encompasses support for clinical trials that use neuroimaging biomarkers as primary or secondary outcome measures, as well as the assembly of patient data into databases, development of tools to analyze these data and their application to addressing clinical translational research and treatment questions. The premise of clinical neuroinformatics is that valuable information can be mined from the vast volume of clinical research and treatment information which is now collected, including diagnostic studies and neuroimaging studies performed for a wide variety of disorders and purposes, if that data could be appropriately transmitted to the scientific community. Since much patient-oriented data are collected in a systematic and standard manner, large databases with de-identification may be an appropriate analytical approach. The strong federal impetus to advance medical informatics is galvanizing efforts to assemble patient-oriented data into such databases, helping to develop solutions to address the critical need for the protection of patient privacy. Rigorous procedures for proper de-identification must be used, since most patient data is sensitive and when initially acquired, completely linked to the patient's identify; alternatively, robust methods to protect against inadvertent disclosure must be determined. Many areas of research in neuroinformatics critical to advancement of this field ranging from development of robust solutions for multisite data management to calibration of cross-site image acquisition are highlighted.

The journal **Neuroinformatics** publishes original articles and reviews with an emphasis on data structure and software tools related to analysis, modeling, integration, and sharing in all areas of neuroscience research. Coverage extends to theory and methodology, including discussions on ontologies or structure of data, modeling approaches, database design, and meta-analyses. Further, there is interest in descriptions of developed databases and software tools, and of the methods for their distribution, massive data sets, and computational simulations of models integrating and organizing complex data (Morse 2008).

As part of this overall goal, we would like to highlight the newly developing field of clinical neuroinformatics, that derives from a large volume of medical imaging and other patient data, acquired during both clinical care of patients as well as clinical research studies. Large collections of neurological data from cohorts of healthy and diseased patients are now currently available from publicly shared databases [e.g. EEG databases (https://epilepsy.uni-freiburg.de; Hunter et al. 2005), the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the OASIS databases of structural MRI data (Jack et al. 2008; Weiner et al. 2010; Marcus et al. 2007; 2010)]. However, in many of these databases the data [i.e., brain magnetic resonance imaging (MRI) or computed tomography (CT)] were acquired specifically for particular

R. L. Gollub
Departments of Psychiatry and Radiology,
Massachusetts General Hospital,
Charlestown, MA 02129, USA

D. A. Turner
Neurosurgery, Neurobiology and Center for Neuroengineering,
Duke University Medical Center,
Durham, NC 27710, USA

R. L. Gollub (✉)
Psychiatric Neuroimaging Research Program,
Massachusetts General Hospital,
Building 120, Suite 100,
Charlestown, MA 02129-2000, USA
e-mail: rgollub@partners.org

research purposes, with the requisite cross-site validation and calibration efforts, to enable coherent assembly of the data across a particular class of patients. Another example is a tightly-defined database of neuroimaging and physiological data, designed to be helpful for planning an individual surgical intervention on patients (D'Hasese et al. 2010). These well-defined data sets are critical for the development of imaging biomarkers that will aid in assessing treatment response, and making prognostic and diagnostic assessments of future patients.

In contrast, most patient data acquired for clinical purposes, while in a standard file format (i.e., DICOM for radiological imaging studies such as brain MRI or CT scans), is not acquired with a single standardized image acquisition protocol nor are the patient's clinical metadata standardized to capture quantitative metrics of disease severity and/or neuropsychological functioning. However, the immense quantity of such clinical data acquired on a daily basis may be very helpful, for example to identify what "normal" imaging studies really include, what is on the border of pathological and when studies clearly show true pathological findings. Unfortunately, the current state of the art of clinical studies (particularly neuroimaging but also including other modalities) has a wide range of "normal", and many normal variants are often described as pathological, since they may not be frequently encountered. Future applications of such large clinical collections of structural and functional imaging data include the potential to map out patterns of healthy development, normal variants (i.e., location of speech and motor areas, for example) or to infer the borders of pathological entities. From another perspective, these collections of standardized databases of medical data (with associated clinical metadata) could be used to improve radiological assessment of routine imaging data. For example, radiologists often use numeric rating scales to quantify severity of pathology (e.g. calcification or bone loss) observed in research and clinical imaging scans. Development of robust, sensitive clinical rating scales would be greatly facilitated by providing a large number of radiologists with access to large databases of well-characterized medical image datasets so that their assessments could be used to generate metrics of reliability and generalizability for a rating scale.

Other examples of such large databases in development include genomic data, which may eventually be useful to mapping out personal drug sensitivity and drug resistance, as well as the tendency to develop disease, once sufficiently large samples have been acquired to truly sort out normal variants from pathological tendencies (Saito et al. 2009; Stenson et al. 2009). Existing large databases of clinical data include the National Inpatient Sample (NIS; Patil et al. 2005) and other representative databases of clinical data derived from ordinary clinical care. A large database of genomic information acquired from glioblastoma tumor samples has led to highly novel conclusions about mechanisms of tumor growth (Parsons et al. 2008).

Neuroinformatics focuses on developing databases to allow and facilitate these goals, as well as on the development of tools to derive meaning from such archived data. A current and highly relevant issue is to how to assemble meaningful databases of similar datasets, so that the conclusions arising from the data are sufficiently valid to apply to a wide range of clinical data. Enormous efforts have gone into the standardization of quantitative medical informatics such as that for human clinical assessments and laboratory values (i.e., blood counts, body chemistry, LOINC and HL7) (Forrey et al. 1996; Dolin et al. 2001). These efforts have enabled the development of "normal" values and have a tremendous impact on clinical decision-making and support. There are now efforts underway to use these clinically collected data collections for clinical translational research (Murphy et al. 2010).

Neuroinformatics Goals for Clinical Data

Clinical neuroinformatics encompasses multiple purposes for both developing databases and tools for accessing and analysis of the underlying data, for various purposes—the sheer scale of the databases now requires advanced tools for access and for deriving meaning and knowledge.

In this sense there are two goals of assembling and analyzing clinical data:

1) The first goal is to identify common patterns across patients that may represent pathological tendencies, sufficient to eventually confirm a specific diagnosis (i.e., functional MRI brain patterns associated with either psychotic symptoms or the diagnosis of schizophrenia) (Roffman et al. 2008; White et al. 2009). This goal requires the removal of individual patient variations (i.e., size of the brain and skull) and reformatting clinical data into a common, uniform pattern, such as an atlas or standard view, so that pathological variations can be recognizable when removed from the individual patient context;
2) The second goal is then to take those patterns, warp them back to an individual patient's morphology or function, and see if abnormalities apply to a specific patient, i.e. to help with diagnosis or etiology of the disease, or to treat a specific disease (D'Hasese et al. 2010). For example, this may involve taking a standard brain view or atlas and warping to an individual patient scan to identify a specific region of the brain for treatment targeting.

Patient data now includes a wide range of quantitative values, usually acquired for the patient's diagnosis and treatment. These values include ordinary laboratory (i.e., blood derived) data such as blood counts and blood chemistry, pathological data derived from tissues or fluids, a wide variety of imaging data acquired from throughout the body, and secondary physiological data, such as electroencephalogram (EEG) or electromyogram (EMG) data. Typical hospital databases also commonly include electronic medical records (i.e., consultation and clinic reports, operative reports) and imaging reports, and sometimes, actual imaging data. Imaging records include original digital scans (such as MRI and CT scans), angiography (including computed tomographic angiographic studies), radio nucleotide scans (i.e., bone scans, PET scans), ultrasound, and other modalities. In many cases, the hospital databases specifically exclude searching for specific abnormalities across a wide range of patient data or more than one patient for confidentiality reasons, and focus instead purely on clinical treatment issues, one patient at a time, to avoid issues associated with patient privacy and confidentiality as much as possible.

One of the main problems with clinical data is that linked metadata on the patient's medical condition is very valuable, but one cannot keep linked metadata and at the same time preserve anonymity, so that patient identification and confidentiality is a real and pervasive issue. There is a need to have several paths to ensure that patients are not specifically being named, but on the other hand to be able to maintain as much of the linked metadata as possible. For example, summarized medical data are now available in generic databases, such as the National Inpatient Sample (NIS), which are fully de-identified and provide a representative view of inpatient encounters in the US (Patil et al. 2005). However, one cannot link encounters, so that if a patient is readmitted for a secondary problem (such as an infection) then the patient cannot be linked to their primary admission, to see the timing and occurrence of the infection after a specific procedure, for example.

Rationale to Develop Databases and Analysis Tools

There are several reasons to invest the considerable time and expense to construct an archive of clinical data, and particularly to provide this to other investigators or clinicians in an open, understandable and usable format. Examples include:

1) Quality control and quality improvement: Many academic hospitals now share quality control data for a wide range of diagnoses and procedures, to provide benchmarks for patient outcomes, such as infections and mortality, usually in relation to patient acuity or disease processes.
2) Direct patient care: Most hospitals now have comprehensive internal databases and some are starting to allow data mining in support of improvements in patient care. These large patient care databases are extensive, usually including all labs, pathology, radiological studies and are typically structured to optimally organize information for individual patients. Goals for these efforts include improved communication across treatment teams and continuity of care over time. The format of these databases often precludes research, but search strategies may eventually be cautiously allowed.
3) Clinical research: Robust software solutions for securely transferring large volumes of data from hospital internal databases into research warehouses enables clinical scientists to use this information for secondary research purposes (i.e. Murphy et al 2010). Such repositories are invaluable for identifying unique cohorts of subjects for recruitment into prospective research studies, for answering research questions that require large cohorts of patient data to detect subtle findings or rare variants. This research may be able to help differentiate normal variants from pathological processes, for example, particularly when patient continuity across years of follow-up can be performed.
4) Clinical Trials: There are a number of solutions for linking the key constituents of these studies including clinician scientists, point of care (i.e. hospital or clinic), imaging facility, and study sponsors (i.e. government, pharmaceutical or biotechnology companies, foundations). There remain many challenges to efficient conduct and robust implementation and subsequent data mining of the collected data.
5) Derived clinical information: A number of existing, large databases have been developed for understanding clinical care at a large scale, including NIS (National Inpatient Survey), Medicare databases, and VA care databases—these have clinical codes (i.e., ICD-9) for all admissions, procedures, complications, and are excellent resources for deriving long-term trends in clinical care and complications (i.e., Patil et al 2005).
6) Patient care and outcome databases: These involve billing, outcome, mortality, now routinely shared across consortia of medical centers, often for cost comparison, superficial outcomes (i.e., death, infections, etc.)
7) Disease genomics databases: This would include data on known genetic mutations, for example, as extensions of disease analysis. A recent example is a wide screen of known genetic data for brain tumors, identifying previously unrecognized genes and functional proteins

as possibly being important in brain tumor formation (IDH2; Parsons et al. 2008).

8) Personalized genomics: Genomic medicine or individualized medicine is now a hot topic—the goal is to obtain large-scale array (10–20 K genes) on everyone to 1) identify their susceptibility to disease (cardiac, diabetes, breast cancer, neurofibromatosis, degenerative diseases, etc.), 2) to assess response to medications and dose, longevity, and 3) to "personalize" their treatments/interventions to maximize response (Saito et al. 2009). Privacy is a large issue, since the insurance companies would also like to have this information to maximize their profits and to eliminate high-risk patients. If one had outcome as well as the full-scale genetic information then data mining for disease intervention would be critical, and could be used to develop correlations towards outcomes.
9) Development of "standard" patients: Imaging databases tend to focus on placing individual patients into a common coordinate system, by warping into a "standard" brain, so that individuals can be easily compared for research purposes. However, for clinical purposes (i.e., brain surgery), a database should be warped and mapped onto an individual brain to help decide possible functional localization and variation between brains (some older atlases included variation and statistical measures of differences of structures). An example of this form of database, developed for individual surgical intervention is CranialVault (D'Hasese et al. 2010).
10) Brain-Machine Interface: The implementation of brain-machine interface technology requires large-scale processing of multiple sites of information (in real time) for assessing "state" of circuits in the brain (Nicolelis and Lebedev 2009). This technology is highly challenging requiring development of rapid analysis tools, a high-throughput output to external devices and direct implementation in neuroprosthetics.

Summary

Only a small fraction of commonly acquired clinical data is accessible for either analysis or to place in appropriate, de-identified databases for subsequent analysis. Part of the issue is the sheer volume of commonly acquired clinical data and the lack of specific conditions under which the data is collected, hence lowering the usefulness of the data. Most archives now under development are highly specific to a research goal, where the data were acquired in a well—specified manner, to allow the maximum inference on the secondary analysis of the data.

Neuroinformatics Special Issue Articles

The articles in this special issue highlight some of the current approaches to the implementation and use of clinically relevant neuroimaging databases and analysis tools.

The article by Adamson and Wood (2010) presents a suite of tools for managing neuroimaging data, with a scheme for data analysis. The toolkit enables anonymous data retrieval. This suite of tools (DFBIdb) may eventually be freely available to facilitate construction of databases.

A challenging scenario for clinical neuroinformatics is to provide a viable solution when the imaging data must be stored in more than one location. One solution is to provide a "federated" database, which actually provides data storage across multiple, internet-accessible locations but in a common format. The article by Ozyurt et al. (2010) highlights the difficulties in data management for such federated databases and poses one evolving solution developed by the Function Biomedical Research Informatics Network (fBIRN) team. Their software tools provide sufficient flexibility in the database extension to support storage of new types of data after implementation. This is a common problem; the field of neuroimaging is evolving at such a rapid pace that initial conceptualization of a database schema will quickly be outdated so there must sufficient flexibility for continued inclusion of relevant datasets.

Repositories of valuable clinical data and associated neuroimaging data collected by multisite consortia are being made accessible to the scientific community. These include multi-site schizophrenia studies and associated imaging data acquired through the Mind Research Network (http://www.mrn.org/) and BIRN (http://www.birncommunity.org/) that have led to multiple publications (i.e. Potkin et al. 2009; Kim et al. 2009; Roffman et al. 2008; White et al. 2009). The article by Kim et al. (2010) is a novel analysis of extant data within the database, using a combination of extracted fMRI activation indices and summary scores from neuropsychological tests, to attempt to find feature combinations that might provide biological insight into the complex disorder known as schizophrenia. The authors conclude that a widespread frontal network might subserve some of the abnormalities noted in the disorder. This study demonstrates some of the important implications of developing a database, where additional users may provide analysis or insights not originally envisioned at the time the study was designed.

Many of the critical challenges faced by key stakeholders in clinical trials arise from the necessity of agreeing upon a common data archive to store both clinical information as well as imaging data, which requires a common format of original data and a method of managing workflow across the various sites and central data repository. The article by El-Ghatta et al. (2010) provides a cogent articulation of the challenges, and most importantly, offers a viable, shared, open source software solution to address the workflow management issue based on the use of grid computing and the caGrid architecture (http://cagrid.org/). This software allows clear interoperability across the various sites often included within the clinical trial.

Though the above databases are derived from direct patient data or clinical research, other domains of neuroinformatics include the development of secondary databases. For example, the article by Zaveri et al. (2010) describes an approach to quantitatively assessing the quality of the medical literature in a medical evidence format, by encoding specific features of randomized trials (amongst others). This quantitative approach may provide a simpler solution to estimating the clinical worth of the available literature on a topic, by providing categories of information provided in the articles, as well as a common database of such articles. This approach may facilitate, for example, creating meta-analysis of existing articles, as well as formulating guidelines based on published data.

Conclusions

Clinical neuroinformatics is poised to take advantage of the large explosion of electronically captured, clinically derived data, both from ordinary clinical practice as well as from research studies and clinical trials. There are many possible approaches to analyzing and summarizing such data, both across multiple patients to infer disease states and disease mechanisms, as well as assessing how well any one patient may or may not show pathological abnormalities. Databasing and analysis tools inherent to neuroinformatics may provide a solid foundation for the multiple goals inherent in the use of such patient-derived information.

References

Adamson, C. L., & Wood, A. G. (2010). DFBIdb: A software package for neuroimaging data management. *Neuroinformatics*. doi:10.1007/s12021-010-9080-z

D'Hasese, P. F., Pallavaram, S., Li, R., Remple, M. S., Kao, C., Neimat, J. S., et al. (2010). CranialVault and its CRAVE tools: a clinical computer assistance system for deep brain stimulation (DBS) therapy. Medical Image Analysis. [Epub ahead of print].

Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., et al. (2001). The HL7 clinical document architecture. *Journal of the American Medical Informatics Association, 8*, 552–569. PMID: 11687563.

El-Ghatta, S. B., Cladé, T., & Snyder, J. C. (2010). Integrating clinical trial imaging data resources using service-oriented architecture and grid. *Neuroinformatics*. doi:10.1007/s12021-010-9072-z

Forrey, A. W., McDonald, C. J., DeMoor, G., Huff, S. M., Leavelle, D., Leland, D., et al. (1996). Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry, 42*, 81–90. PMID: 8565239.

Hunter, M., Smith, R. L. L., Hyslop, W., Rosso, O. A., Gerlach, R., Rosta, J. A. P., et al. (2005). The Australian EEG database. *Clinical EEG and Neuroscience, 36*, 76–81.

Jack, C. R., Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging, 27*, 685–691. PMID: 18302232.

Kim, D., Manoach, D., Turner, J., Mannell, M., Brown, G., Ford, J., et al. (2009). Dysregulation of working memory and default-mode networks in schizophrenia during a Sternberg item recognition paradigm. *Human Brain Mapping, 30*, 3795–3811.

Kim, D., Sui, J., Rachakonda, S., White, T., Manoach, D. S., Clark, V. P., et al. (2010). Identification of imaging biomarkers in schizophrenia: a coefficient-constrained independent component analysis of the Mind multi-site schizophrenia study. *Neuroinformatics*. doi:10.1007/s12021-010-9077-7.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience, 19*, 1498–1507.

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience, 22*, 2677–2684.

Morse, T. M. (2008). Neuroinformatics: from bioinformatics to databasing the brain. *Bioinformatics and Biology Insights, 2*, 259.

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., et al. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Information Association, 17*, 124–130. PMID: 20190053.

Nicolelis, M. A. L., & Lebedev, M. A. (2009). Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nature Reviews. Neuroscience, 10*, 530–540.

Ozyurt, I. B., Keator, D., Wei, D., Fennema-Notestine, C., Pease, K., Bockholt, J., et al. (2010). Federated web-accessible clinical data management within an extensible neuroimaging database. *Neuroinformatics*. doi:10.1007/s12021-010-9078-6.

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science, 321*, 1807–1812.

Patil, P. G., Pietrobon, R., & Turner, D. A. (2005). National trends in surgical procedures for degenerative cervical spine disease: 1990–2000. *Neurosurgery, 57*, 753–758.

Potkin, S. G., Turner, J. A., Brown, G. G., McCarthy, G., Greve, D. N., Glover, G. H., et al. (2009). Working memory and DLPFC

inefficiency in schizophrenia: The FBIRN study. *Schizophrenia Bulletin, 35*, 19–31. PMID: 19042912.

Roffman, J. L., Gollub, R. L., Calhoun, V. D., Wassink, T. H., Weiss, A. P., et al. (2008). MTHFR 677C→T genotype disrupts prefrontal and dopaminergic function in schizophrenia. *Proceedings of the National Academy of Science U S A, 105*, 17573–17578. PMCID: PMC2582272.

Saito, T. L., Yoshimura, J., Sasaki, S., Ahsan, B., Sasaki, A., Kuroshu, R., et al. (2009). UTGB toolkit for personalized genome browsers. *Bioinformatics, 25*, 1856–1861.

Stenson, P. D., Mort, M., Ball, E., Howells, K., Phillips, A. D., et al. (2009). The human gene mutation database: 2008 update. *Genome Medicine, 1*, 13.

Weiner, M. W., Aisen, P. S., Jack, C. R., Jr., Jagust, W. J., Trojanowski, J. Q., Shaw, L., et al. (2010). The Alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia, 6*(3), 202–11.e7. PMID: 20451868.

White, T., Magnotta, V., Bockholt, H. J., Williams, S., Gollub, R. L., Mueller, B. A., et al. (2009). Global White matter abnormalities in schizophrenia: a multicenter diffusion tensor imaging study. Schizophrenia Bulletin. [Epub ahead of print]. PMID: 19770491.

Zaveri, A., Cofiel, L., Shah, J., Pradhan, S., Chan, E., Dameron, O., et al. (2010). Center for Excellence in Research Reporting in Neurosurgery (CERR-N)—achieving high quality. *Neuroinformatics*, in press (NEIN-D-09-00028).