NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The image and data archive at the Laboratory of Neuro Imaging

Karen L. Crawford, Scott C. Neu, Arthur W. Toga *

Laboratory of Neuro Imaging, USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Accepted 27 April 2015
Available online xxxx

Keywords:
IDA
Data repository
Data sharing

ABSTRACT

The LONI Image and Data Archive (IDA)[1] is a repository for sharing and long-term preservation of neuroimaging and biomedical research data. Originally designed to archive strictly medical image files, the IDA has evolved over the last ten years and now encompasses the storage and dissemination of neuroimaging, clinical, biospecimen, and genetic data. In this article, we report upon the genesis of the IDA and how it currently securely manages data and protects data ownership.

Introduction

The IDA was initially created to de-identify and collect neuroimaging data for the International Consortium for Brain Mapping (ICBM) study in which Magnetic Resonance Imaging (MRI) and Positron-Emission Tomography (PET) scans from 850 normal adult subjects were collected at three North American sites (Mazziotta et al., 1995, 2009; Kochunov et al., 2002). As interest in storing data in biomedical repositories grew, the number and size of studies utilizing the IDA expanded considerably. The IDA has become a global resource for storing and disseminating neuroimaging, clinical, biospecimen, and genetic data for a growing number of national and international consortia efforts and many smaller, single-center studies.

Background

Early IDA development proceeded in concert with new HIPAA (Health Insurance Portability and Accountability Act (US Department of Health and Human Services)) regulations that took effect in 2003, which put emphasis on maintaining patient confidentiality throughout data collection and data collaboration activities. Over time, attention on data sharing dynamics (Amari et al., 2002; Eckersley et al., 2003; Gardner et al., 2003; Koslow, 2000; Kulynych, 2002; Toga and Dinov, 2015) grew in parallel with the launch of increasingly complex multi-site consortia studies. The number, scale and scope of collaborations utilizing the IDA expanded and extended beyond data acquisition sites to involve organizations performing quality assessments and to external electronic data capture (EDC) systems. Consequently, IDA functionality and features were expanded to accommodate the specific needs of different types of studies and users.

Today the IDA contains both raw (direct from scanner) and processed (output from processing programs) neuroimaging data, clinical data, and analysis results for dozens of studies on Alzheimer's disease (Toga and Crawford, 2010), multiple sclerosis, Huntington's disease, Parkinson's disease (Marek et al., 2011), traumatic brain injury, normal development, HIV/AIDS, bipolar disorder, schizophrenia, and others. Sites across North America, Europe, Australia, and Asia have been actively uploading data since 2003 with the average number of newly added images reaching over 5000 per month in 2014. For a growing number of studies, clinical data, image quality assessments, and analysis results are uploaded to the IDA on a daily basis. For many studies the IDA is the exclusive location for pooled data but for a small number of studies (Australian Imaging, Biomarkers and Lifestyle (Ellis et al., 2009), Autism Brain Imaging Data Exchange (Di Martino et al., 2014), Brain Genomics Superstruct Project (Buckner et al., 2012), and Human Connectome Project (Rosen et al., 2010)) the IDA mirrors data available in other repositories and this allows users to obtain data from all these studies within a single system. Worldwide, the number of image downloads from the IDA exceeds 7 million (Fig. 1). The logo for each study, along with a link to the study's web site, appears at the top of each IDA web page in order to focus on the study rather than the IDA. There is no requirement to acknowledge the IDA in publications that use data obtained through the IDA.

Data collection

The IDA presently holds data from more than 70 studies and 125 different institutions, and is continually receiving new data. Table 1 lists a subset of research studies that are storing data in the IDA. On average, more than 120 raw scans are uploaded each weekday from sites located in the Americas, Europe, and Australia. There are currently over 350,000 neuroimaging scans (over 96 million files) archived in the IDA of which

* Corresponding author at: Laboratory of Neuro Imaging, Keck School of Medicine of USC, 2001 N. Soto Street, SSB1-102, Los Angeles, CA 90032, USA.
E-mail address: toga@loni.usc.edu (A.W. Toga).

[1] https://ida.loni.usc.edu/.

http://dx.doi.org/10.1016/j.neuroimage.2015.04.067
1053-8119/

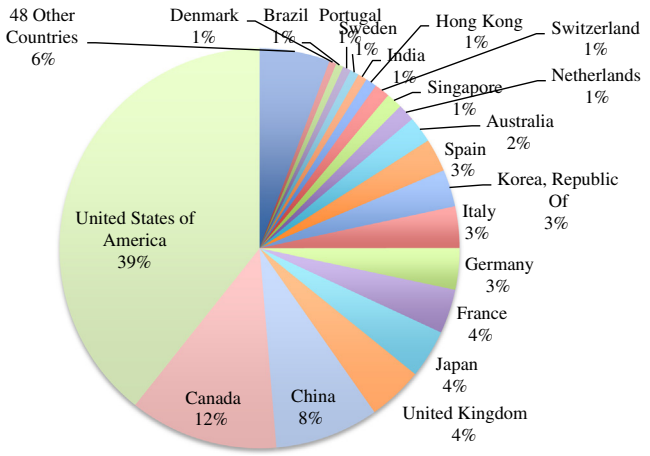

Fig. 1. Users in 68 countries have downloaded over 7 million raw and processed images from the IDA since 2004.

64% are raw and 36% are processed scans. These scans consist of structural MRI, functional MRI (fMRI), diffusion MRI, magnetic resonance angiography (MRA), positron emission tomography (PET), computed tomography (CT), and single-photon emission computed tomography (SPECT) data from tens of thousands of human subjects (some followed longitudinally for over a decade) and hundreds of phantom scans used for image quality control.

User interactions occur primarily through web-browsers that incorporate the Java[2] plugin. Since minimal technical proficiency is required, new studies can quickly come online and begin to upload data without requiring lengthy training or devoting excessive time and resources at the participating sites.

Uploading and de-identifying data

Users can upload neuroimaging data files in many formats, including those listed in Table 2. During the data archiving process, the user opens a web-browser and logs into the web application. The web application deploys a Java applet that runs on the user's computer at the acquisition site. The applet automatically detects image file formats and invokes format-specific de-identification programs to remove patient-identifying information from the files before transferring the de-identified files to the IDA. Different de-identifications are used for different scanner types and different file formats in accordance with the individual needs of the studies that are involved. De-identification programs may be customized for the needs of the study, however the general approach involves the replacement of patient name and patient ID fields with the user-supplied research identifier, removal of all elements that are either not of a preserved type (e.g. numeric, code string) or in a set of preserved elements, and hashing of elements containing unique identifiers. Several scanner-specific private elements are also retained in order to preserve needed information, for example gradient information for diffusion scans. For fMRI scans, a paradigm file may be uploaded as an attachment. The fMRI attachment, which is linked to the fMRI image, will be provided whenever a user downloads the fMRI image files. Processed neuroimaging files are uploaded in conjunction with a structured processing provenance file. The provenance file contains information about the image processing workflow and identifies the image(s) from which the processed image was derived. A schema enforces the structure and content of the provenance files and the subject and image identifiers are validated at time of upload. Once the de-identified files are uploaded into the IDA, metadata from the image file headers (and provenance files) are extracted and used to detect duplicate data and classify images (Neu et al., 2012) into subtypes (e.g., structural MRI, functional MRI, or diffusion MRI). This metadata is then stored in the database and combined with other information from the upload to support future queries on the data.

Uploaders may also send clinical data and analysis results to the IDA using a tool that transfers data in comma-separated value (CSV) files. Validity checks are performed on all CSV files before they are accepted to help ensure data quality. Many studies use the tool to copy the entire contents of clinical EDC systems to the IDA on a daily basis. The data transfer tool supports both incremental updates and full synchronization of entire data sets. Once this clinical data is made available, investigators may access all image and clinical data from a study through the IDA. This frees the EDC systems, which are focused on data collection, from having to manage access to the data.

Table 1
Representative research studies utilizing the IDA.

Study name	Centers	Subjects	Deposit activity	Downloads[a]
Aging & dementia				
Alzheimer's Disease Neuroimaging Initiative (ADNI)	58	2469	2005–present	7,400,000
Australian Imaging, Biomarkers and Lifestyle (AIBL)	1	810	2008–present	68,000
Imaging & Genetic Biomarkers for AD	1	177	2009–present	Private
HIV				
Age Moderates HIV-Related CNS Dysfunction	1	116	2006–2007	Private
Cardiovascular & HIV/AIDS Effects on Brain & Cognition	4	349	2009–2014	Private
Huntington's disease				
Huntington's Disease Neuro Imaging Initiative (HDNI)	4	369	2007–2011	Private
Track-On Huntington's Disease	4	242	2012–2014	Private
Brain injury				
Effects of TBI & PTSD on Alzheimer's Disease in Vietnam Vets (DoD ADNI)	17	115	2013–present	15,000
Volumetrics in Brain Trauma	1	393	2006–present	Private
Transforming Research and Clinical Knowledge in TBI (Track-TBI)	11	418	2014–present	7000
Normal development				
International Consortium for Brain Mapping (ICBM)	3	852	2003–2009	123,000
Genetic influences on the brain: A twin study	1	1045	2007–2013	Private
Multiple sclerosis				
Hippocampal Volume Loss in Multiple Sclerosis	1	58	2007–2010	Private
Multi-center Estriol Study	17	334	2007–2014	Private
Parkinson's				
Parkinson's Progression Markers Initiative (PPMI)	31	1230	2011–present	360,000
Hippocampal atrophy in Parkinson's disease	1	166	2008–2011	Private
Schizophrenia				
North American Prodrome Longitudinal Study (NAPLS)	8	845	2009–present	Private
Studies mirrored in the IDA				
Autism Brain Imaging Data Exchange (ABIDE)			2012	48,000
Brain Genomics Superstruct Project (GSP)	1	1570	2014	50

[a] Indicates the number of images, clinical and genetic data sets downloaded from the IDA for studies with open data sharing policies.

[2] http://www.java.com/en/.