Recommendations for Processing Head CT Data

John Muschelli[a]

[a]*Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, 615 N Wolfe St, Baltimore, MD, 21205*
[b]*Johns Hopkins Hospital, Department of Neurology, 601 N Caroline St, Baltimore, MD 21205*
[c]*Brain Injury Outcomes, Johns Hopkins University, 750 East Pratt Street, Baltimore, MD 21202*

Abstract

This is the abstract.

It consists of two paragraphs.

Introduction

Many research applications of neuroimaging use magnetic resonance imaging (MRI). MRI allows researchers to study a multitude of applications and diseases, including studying healthy controls. Clinical imaging, however, relies heavily on X-ray computed tomography (CT) scans for diagnosis and prognosis. Studies using CT scans cannot generally recruit healthy controls or large non-clinical populations due to the radiation exposure and lack of substantial benefit. As such, much of head CT data is gathered from prospective clinical trials or retrsopective studies based on health medical record data and hospital picture archiving and communication system (PACS). We wish to provide some recommendations and guidelines from our experience with CT data, as well as insights from working with MRI stuies. We will discuss existing software options for neuroimaging in general and those that are specific to CT throughout the paper.

We will focus on aspects of quantitatively analyzing the CT data and getting the data into a format familiar to most MRI neuroimaging researchers. Therefore, we will not go into detail of DICOM reading tools or imaging suites for radiologists, which are generally proprietary and quite costly. Moreover, we will be focussing specifically on non-contrast head CT data, though many of the recommendations and software is applicable to images of other areas of the body.

Data Organization

Most of the data coming from a PACS is in DICOM (Digital Imaging and Communications in Medicine) format. Generally, DICOM files are a combination of metadata about the image (also called a header) and the individual pixel data, many times embedded in a JPEG format. The header has a collection of

Email address: jmusche1@jhu.edu (John Muschelli)

information, usually referred to as fields or tags. Tags are usually defined by a set of 2 hexadecimal numbers, which are embedded as 4 alphanumeric characters. For example, `(0008,103E)` denotes the `SeriesDescription` tag for a DICOM file. Most DICOM readers extract and use these tags for filtering and organizing the files.

We will use the phrase scanning session (as opposed to "study" and reserve study to denote a trial or analysis), a series for an individual scan, and a slice for an individual picture of the brain. Each series (`Series Instance UID` tag) and scanning session (`Study Instance UID` tag) should have a unique value in the DICOM header that allows DICOM readers to organize the data by scanning session and series.

DICOM Anonymization

One of the common issues with DICOM data is that a large amount of protected health information (PHI) can be contained in the header. DICOM is a standard where individual fields in the header are to contain the same values across different scanners and sites, but only if that manufacturer and site are diligent to ascribing to the DICOM standard. Though many DICOM header fields are consistent across neuroimaging studies, duplicate data and additional fields may be required to obtain the full amount of data required for analysis. Moreover, different scanning manufacturers can embed information in non-standard fields. The goal is to remove these fields if they contain PHI but retain these fields if they embed relevant information of the scan for analysis. These fields then represent a challenge to a "lossless" anonymization if the data do not conform to a standard across scanning sites, manufacturers, or protocols.

We will discuss reading in DICOM data and DICOM header fields in the next section. Though thse steps can be crucial for extracting information from the data, many times the data must be shared or transferred before analysis. Depending on the parties receiving the data, anonymization of the data must be done first. Aryanto, Oudkerk, and Ooijen (2015) provides a look at a multitude of options for DICOM anonymization and recommends the RSNA MIRC Clinical Trials Processor (CTP, https://www.rsna.org/research/imaging-research-tools) cross-platform Java software as well as the DICOM library (https://www.dicomlibrary.com/) upload service. We also recommend the DicomCleaner cross-platform Java program as it has fit many of our needs. Bespoke solutions can be generated using `dcm4che` (such as `dcm4che-deident`) and other DICOM reading tools (discussed below), but many of these tools have built-in capabilities that are difficult to add (such as removing PHI embedded in the pixel data).

Reading DICOM data

We will focus on 2 analysis platforms for statistical analysis, including `R` (CITE) and `Python` as well as standalone software. Other imaging platforms such as the Insight Segmentation and Registration Toolkit (ITK) are great pieces of software that can perform many of the operations that we will be discussing. Moreover, `MATLAB` has an extensive general imaging suite, as well

as large neuroimaging platforms such as SPM (CITE). We will touch on some of this software with varying levels. We aim to present software that we have had used directly for analysis or preprocessing. Also, other papers and tutorials discuss their use (CITE).

For reading DICOM data, there

Converting DICOM to NIfTI

Though most imaging analysis tools can read in DICOM data, there are downsides to using the DICOM format. In most cases, a DICOM image is split into a series of slices, where each slice is a different file. This separation can be cumbersome on data organization if using folder structures. As noted above, these files also can contain a large amount of PHI. Some formats may be compressed using proprietary compression such as JPEG2000; alternatively, if data are not compressed file storage is inefficient. Most importantly though, many imaging analyses perform 3-dimensional operations, such as smoothing. Thus, putting the data into a different format may be helpful.

Many different general 3D medical imaging formats exist, such as ANALYZE, NIfTI, NRRD, and MNC. We recommend the NIfTI format as it can be read by nearly all medical imaging platforms, has been widely used, has a format standard, can be stored in a compressed format, and is how much of the data is released online. Moreover, we will present specific software to convert DICOM data and the recommended software (`dcm2niix`) outputs data in a NIfTI file.

Although we recommend this software, many good and complete solutions exist. Examples include `dicom2nifti` in the `read.dicom` `R` package, `pydicom`, `dicom2nifti` in `MATLAB`, and using large imaging suites such as using `ITK` image reading functions for DICOM files and can write NIfTI outputs. We recommend the `dcm2niix` (LINK) function from Chris Rorden for CT data. The reasons are 1) it works with all major scanners, 2) incorporates gantry-tilt correction for CT data, 3) can handle variable slice thickness, 4) is open-source, 5) is fast, 6) has responsive developers, and 7) works on all 3 major operating systems (Linux/OSX/Windows) (TRUE??). Moreover, the popular AFNI neuroimaging suite includes a `dcm2niix` program with its distribution. In `R`, the `divest` package (LINK) and the `XXXXXXXX` Python module wraps the underlying code for `dcm2niix` to provide the same functionality of `dcm2niix`, along with the ability to manipulate and subset the header data as necessary.

We will describe a few of the features above. In some head CT scans, the gantry is tilted to reduce radiation exposure to non-brain areas, such as the eyes. Thus, the slices of the image are at an oblique angle. If slice-by-slice analyses an affine registration (as this tilting is a shearing) are done, this tilting is not an issue. This tilting does cause issues for 3D operations as the distance of the voxels between slices is not correct and especially can show odd visualizations (FIGURE). The `dcm2niix` output returns both the corrected and non-corrected image. As the correction moves the slices to a different area, `dcm2niix` may pad the image so that the entire head is still inside the field of view. As such, this may cause issues with algorithms that require the 512x512 axial slice dimensions. Though less common, variable slice thickness can occur in reconstructions where

only a specific area of the head is of interest. For example, an image may have 5mm slice thicknesses throughout the image, except for areas near the third ventricle, which has a 2.5mm slice thickness. To correct for this, `dcm2niix` interpolates between slices to ensure each image has a consistent voxel size. Again, `dcm2niix` returns both the corrected and non-corrected image.

Once converted to NIfTI format, one should ensure the scale of the data. Most CT data is betweeen -1024 and 3071 HU. Values less than -1024 are commonly found due to areas of the image outside the field of view that were not actually imaged. One first processing step would be to Winsorize the data to the $[-1024, 3071]$ range. After this step, the header elements `scl_slope` and `scl_inter` elements of the NIfTI image should be set to 1 and 0, respectively, to ensure no data rescaling is done in other software. Though HU is the standard format used in CT analysis, negative HU values may causes issues with standard imaging pipelines built for MRI, which typically have positive values. Rorden (CITE) proposed a lossless transformation, called Cormack units, which have a minimum value of 0.

Brain Extraction in CT

Head CT data typically contains the subject's head, face, and maybe neck and other lower structures, depending on the field of view. Additionally, other artifacts are typically present, such as the pillow the subject's head was on, the bed/girney, and any instruments in the field of view. We do not provide a general frameowrk to extract the subject from the artifact data, but provide some recommendations for working heuristics. Typically the range of data for a subject is within -100 to 300, excluding the skull, other bones, and calcificiations. Creating a mask from this data range tends to remove the bed/girney, most intstruments, the pillow, and the background. Retaining the largest connected component, filling holes (to include the skull), and masking the original data with this resulting mask will return the subject. Note, care must be taken whenever a masking procedure is used with HU values as 0 is a real value: if all values are set to 0 outside the mask in an image with HU values, the value of 0 corresponds to both 0 HU and outside of mask. Either transforming the data into Cormack units, adding a value to the data (such as 1025) then setting values to 0, or using `NaN` are recommended in negative values are of interest.

One of the most common steps in processing imaging of the brain is to remove non-brain structures from the image.

Registration to a CT template

Conclusions

Most

- DICOM data

  - DICOM anonymizer
  - longitudinal data
    * keep time, but do days from Jan-01-1900
    * or could keep date
- Remove Localizers and scouts

Skull Stripping

-

- clinical toolbox
-

References

Aryanto, KYE, M Oudkerk, and PMA van Ooijen. 2015. "Free DICOM de-Identification Tools in Clinical Research: Functioning and Safety of Patient Privacy." *European Radiology* 25 (12): 3685–95.