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 studies. 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).

A note on de-identification: time between scans

Although most of the presented solutions are good at anonymization and de-identification of the header information, only a few such as CTP, have the utilities required for longitudinal preservation of date differences. Though it can be debated whether dates are identifiable information, some clinical trials and other studies rely on serial CT imaging data, and the differences between times are crucial to determine when events occur.

Reading DICOM data

We will focus on 2 analysis platforms for statistical analysis, including R (CITE) and Python as well as standalone software. The main reasons are that R and Python are free, open source, and we have a working knowledge of the utilities these have. We are also involved in the Neuroconductor project (https://neuroconductor.org/) (Muschelli et al. 2018), which is a repository of R packages for medical image analysis. 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 are multiple options. The MATLAB imaging toolbox, `oro.dicom` R package, `pydicom`, and ITK cinterfaces can read DICOM data amongst others. The DICOM toolkit `dcmtk` has multiple DICOM manipulation tools, including `dcmconv` to convert DICOM files to other imaging formats.

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.

Converting DICOM to NIfTI

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 `oro.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. We have published a method that uses the brain extraction tool (BET) from FSL, originally built for MRI, to perform brain extraction (CITE) with code provided (http://bit.ly/CTBET_BASH). Many papers present brain extracted CT images, but do not always disclose the method of extraction. Recently, convolutional neural networks and shape propagation techniques have been quite successful in this task (Akkus et al. 2018) and models have been released (https://github.com/aqqush/CT_BET). Overall, much research can still be done in this area as conditions such as traumatic brain injury (TBI) and surgery, such as craniotomies or craniectomies can cause these methods to potentially fail.

Registration to a CT template

Though many analyses in clinical data may be subject-specific, population-level analyses are still of interest. In some cases, registration from a template space to a subject space can provide information that can be aggregated across people for analysis. For example, one can perform a label fusion approach to CT data to infer the size of the hippocampus and then analyze hippocampi sizes across the population. One issue with these approaches is that most templates and approaches rely on an MRI template. These templates were developed by taking MRIs of healthy volunteers, which is unethical with CT data due to the radiation exposure. To create templates, retrospective searches through medical records can provide patients who came in with symptoms warranting a CT scan, such as a headache, but had a diagnosis of no pathology or damage. Thus, these neuro-normal scans are similar to that of those collected those in MRI research studies but with some important differences. As these are retrospective, inclusion criteria information may not be easily obtainable if not clinically collected, scanning protocols and parameters may vary, even within hospital and especially over time, and these patients still have neurological symptoms. Though these challenges exist, with a large enough patient population and a research consent at an institution, these scans can be used to create templates and atlases based on CT. To our knowledge, the first publicly available head CT template exists was released in 2012 by Rorden et al. (2012), for the purpose of spatial normalization.

One interesting aspect of CT image registration is again that CT data has units within the same range. To say they are uniformly standardized is a bit too strong in our opinion, but you can think of them as much more standardized than MRI due to the nature of the data. This standardization may warrant or allow the user different search and evaluation cost functions for registration, such as least squares. We have found though that normalized mutual information still performs well in CT-to-CT registration and should be at least considered when using CT-to-MRI or CT-to-PET registration. Along with the template above, Rorden et al. (2012) released the clinical toolbox (https://github.com/neurolabusc/Clinical) for SPM [CITE] to allow researchers to register head CT data to a standard space. However, as the data are in NIfTI format, almost all image registration software should work, though one should consider transforming the units using Cormack units or other transformations as negative values may implicitly be excluded in some software built for MRI registration.

Publicly Available Data

With the issues of PHI above coupled with the fact that most CT data is acquired clinically and not in a research setting, there is a dearth of publicly available data for head CT compared to head MRI. Sites for radiological training such as Radiopedia (https://radiopaedia.org/) have many cases of head CT data, but these are converted from DICOM to standard image formats (e.g. JPEG) so crucial information such as Hounsfield Units and pixel dimensions are lost.

Large repositories of head CT data do exist, though, and many in DICOM format, with varying licenses and uses. The CQ500 (Chilamkurthy et al. 2018) dataset provides approximately 500 head CT scans with different clinical pathologies and diagnoses, with a non-commercial license. The Cancer Imaging Archive (TCIA) has hundreds of CT scans, many cases with brain cancer. TCIA also has a RESTful API, which allows cases to be downloaded in a scripted way; for example, the `TCIApathfinder` R package (Russell 2018) and Python `tciaclient` module provide an interface. The Stroke Imaging Repository Consortium (http://stir.dellmed.utexas.edu/) also has head CT data available for stroke. The National Biomedical Imaging Archive (NBIA, https://imaging.nci.nih.gov) demo provides some head CT data, but are duplicated from TCIA. The NeuroImaging Tools & Resources Collaboratory (NITRC, https://www.nitrc.org/) provides links to many data sets and tools, but no head CT data at this time. The RIRE (Retrospective Image Registration Evaluation, http://www.insight-journal.org/rire/) and MIDAS (http://www.insight-journal.org/midas) projects have a small set of publicly available head CT.

Pipeline

Overall, our recommended pipeline is as follows: 1. Use CTP to organize and anonymize the DICOM data from a PACS. 2. Extract relevant header information for each DICOM, using software such as `dcmdump` from `dcmtk` and store, excluding PHI. 3. Convert DICOM to NIfTI using `dcm2niix`, which can create brain imaging data structure (BIDS) formatted data. Use the tilt-corrected and data with uniform voxel size.

After, depending on the purpose of the analysis, you may do registration then brain extraction, brain extraction then registration, or not do registration at all. If you are doing analysis of the skull, you can also use brain extraction as a first step to identify areas to be removed.

For brain extraction, run `BET` for CT or `CT_BET` (especially if you have GPUs for the neural network).

Concurrent MRI

Additionally, the spatial constrast is much lower than T1-weighted MRI for image segmentation. Therefore, concurrent MRI may be useful. One large issue

Conclusions

We present a simple pipeline for preprocessing of head CT data, along with software options of reading and transforming the data. We have found that though many tools exist for MRI and are applicable to CT data. Noticeable differences exist between the data in large part due to the collection setting (research vs. clinical), data access, data organization, and population-level data. As CT scans provide fast and clinically relevant information and with the increased interest in machine learning in medical imaging data, particularly deep learning using convolutional neural networks, research and quantitative analysis of head CT data is bound to increase. We believe this presents an overview of a useful set of tools and data for research in head CT.

For research using head CT scans to have the level of interest and success as MRI, additional publicly available data needs to be released. We saw the explosion of research in MRI, particularly functional MRI, as additional data were released and and consortia created truly large-scale studies. This collaboration is possible at an individual institution, but requires scans to be released from a clinical population, where consent must be first obtained, and upholding patient privacy must be a top priority. Large internal data sets likely exist, but institutions need incentives to release these data sets. Also, though institutions have large amounts of rich data, general methods and applications require data from multiple institutions as parameters, protocols, and population characteristics can vary widely across institutions.

One of the large hurdles after creating automated analysis tools or supportive tools to help radiologists and clinicians is the reintegration of this information into the healthcare system. We do not present answers to this difficult issue, but note that these tools need to be created to show cases where this reintegration can improve patient care, outcomes, and other performance metrics. We hope the tools and discussion we have provided advances those efforts for researchers starting in this area.

References

Akkus, Zeynettin, Petro M Kostandy, Kenneth A Philbrick, and Bradley J Erickson. 2018. "Extraction of Brain Tissue from CT Head Images Using Fully Convolutional Neural Networks." In *Medical Imaging 2018: Image Processing*, 10574:1057420. International Society for Optics; Photonics.

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.

Chilamkurthy, Sasank, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. 2018. "Deep Learning Algorithms for Detection of Critical Findings in Head Ct Scans: A Retrospective Study." *The Lancet* 392 (10162): 2388–96.

Muschelli, J., A. Gherman, J. P. Fortin, B. Avants, B. Whitcher, J. D. Clayden, B. S. Caffo, and C. M. Crainiceanu. 2018. "Neuroconductor: An R Platform for Medical Imaging Analysis." *Biostatistics*, January.

Rorden, Christopher, Leonardo Bonilha, Julius Fridriksson, Benjamin Bender, and Hans-Otto Karnath. 2012. "Age-Specific Ct and Mri Templates for Spatial Normalization." *Neuroimage* 61 (4): 957–65.

Russell, Pamela. 2018. *TCIApathfinder: Client for the Cancer Imaging Archive Rest Api.* https://CRAN.R-project.org/package=TCIApathfinder.