Recommendations for Processing Head CT Data

John Muschelli [1*]

[1] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA*

Correspondence*:
John Muschelli
jmusche1@jhu.edu

ABSTRACT

Many research applications of neuroimaging use magnetic resonance imaging (MRI). As such, recommendations for image analysis and standardized imaging pipelines exist. Clinical imaging, however, relies heavily on X-ray computed tomography (CT) scans for diagnosis and prognosis. We present tools and recommendations for processing CT data, with emphasis on head CT scans and focusing on open-source solutions. We describe going from raw DICOM data to a spatially normalized brain within CT presenting a full example with code. Overall, we recommend anonymizing data with Clinical Trials Processor, converting DICOM data to NIfTI using 'dcm2niix', using BET for brain extraction, and registration using a publicly-available CT template for analysis.

Keywords: CT, Image Processing

1 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 volunteers. Clinical imaging, however, relies heavily on X-ray computed tomography (CT) scans for diagnosis and prognosis. Studies using CT scans cannot generally recruit healthy volunteers 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 retrospective studies based on health medical record data and hospital picture archiving and communication system (PACS). We discuss transforming this data from clinical to research data and provide some recommendations and guidelines from our experience with CT data similar insights from working with MRI studies. We will discuss existing software options, focusing on open-source tools, 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 imaging suites designed for radiologists, which may be proprietary and quite costly. Moreover, we will be focusing specifically on non-contrast head CT data, though many of the recommendations and software is applicable to images of other areas of the body.

2 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 (i.e. a header) about an image and the individual pixel data, many times embedded in a JPEG format. The header has a collection of 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. The pixel data is usually given in the axial orientation in a high resolution (e.g. 0.5mm^2) grid of 512x512 pixels.

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. The following sections will discuss data organization and data formats.

2.1 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 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, a collection of 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 anonymization without loss of crucial information 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. Reading DICOM data may be necessary for extracting information, but many times the data must be transferred before analysis. Depending on the parties receiving the data, anonymization of the data must be done first. Aryanto et al. (2015) provides a look at a multitude of options for DICOM anonymization and recommend the RSNA MIRC Clinical Trials Processor (CTP, `https://www.rsna.org/research/imaging-research-tools`) a, 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 similar functionality. Bespoke solutions can be generated using `dcm4che` (such as `dcm4che-deident`, `https://www.dcm4che.org/`) 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, aka "burned in").

2.1.1 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. Dates are considered removable identifiable information under HIPAA, some clinical trials and other studies rely on serial CT imaging data, and the differences between times are crucial to determine when events occur or are used in analysis.

2.2   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. All examples in this article use data from 2 subjects within the the CQ500 data set. 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 mostly 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 images 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 small set of publicly available head CT (under 10 participants).

2.2.1   Reading DICOM data

Though `MATLAB` has an extensive general imaging suite, including SPM (Penny et al., 2011), we will focus on R (R Core Team, 2018) `Python` (Python Software Foundation, `https://www.python.org/`), and other standalone software. The main reasons are that R and `Python` are free, open source, and have a lot of functionality with neuroimaging and interface with popular imaging suites. We are also lead 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 well-maintained, useful pieces of software that can perform many of the operations that we will be discussing. 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 the use of these tools in analysis (`https://neuroconductor.org/tutorials`).

For reading DICOM data, there are multiple options. The `oro.dicom` (Whitcher et al., 2011) and `radtools` (Russell and Ghosh, 2019) R packages, `pydicom` Python module (Mason, 2011), `MATLAB` imaging toolbox, and `ITK` (Schroeder et al., 2003) interfaces can read DICOM data amongst others. The DICOM toolkit `dcmtk` (Eichelberg et al., 2004) 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 file is a single slice of the full 3D image series. 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 image data may be compressed, such as JPEG2000 format. Alternatively, if data are not compressed, file storage is inefficient. Most importantly, many imaging analyses perform 3-dimensional (3D) operations, such as smoothing. Thus, putting the data into a different format that handles 3D images as 1 compressed file is desirable. We

110 present examples of reading DICOM data above, but generally recommend using 3D imaging formats and
111 using the above tools to read the DICOM header information.

2.3 Converting DICOM to NIfTI

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

118 Many sufficient and complete solutions exist for DICOM to NIfTI conversion. Examples include
119 `dicom2nifti` in the `oro.dicom` R package, `pydicom`, `dicom2nifti` in `MATLAB`, and using
120 large imaging suites such as using `ITK` image reading functions for DICOM files and then write NIfTI
121 outputs. We recommend `dcm2niix` (`https://github.com/rordenlab/dcm2niix`) (Li et al.,
122 2016) from for CT data for the following reasons: 1) it works with all major scanners, 2) incorporates
123 gantry-tilt correction for CT data, 3) can handle variable slice thickness, 4) is open-source, 5) is fast, 6)
124 is an actively maintained project, and 7) works on all 3 major operating systems (Linux/OSX/Windows).
125 Moreover, the popular AFNI neuroimaging suite includes a `dcm2niix` program with its distribution.
126 Interfaces exist, such as the `dcm2niir` (Muschelli, 2018) package in `R` and `nipype` `Python` module
127 (Gorgolewski et al., 2011). Moreover, the `divest` package (Clayden and Rorden, 2018) wraps the
128 underlying code for `dcm2niix` to provide the same functionality of `dcm2niix`, along with the ability to
129 manipulate the data for more versatility.

130 We will describe a few of the features of `dcm2niix` for CT. In some head CT scans, the gantry is tilted
131 to reduce radiation exposure to non-brain areas, such as the eyes. Thus, the slices of the image are at an
132 oblique angle. If slice-based analyses are done or an affine registration (as this tilting is a shearing) are
133 applied to the 3D data, this tilting may implicitly be corrected. This tilting causes issues for 3D operations
134 as the distance of the voxels between slices is not correct and especially can show odd visualizations
135 (Figure 1A). The `dcm2niix` output returns both the corrected and non-corrected image (Figure 1). As
136 the correction moves the slices to a different area, `dcm2niix` may pad the image so that the entire head
137 is still inside the field of view. As such, this may cause issues with algorithms that require the 512x512
138 axial slice dimensions. Though less common, variable slice thickness can occur in reconstructions where
139 only a specific area of the head is of interest. For example, an image may have 5mm slice thicknesses
140 throughout the image, except for areas near the third ventricle, where slices are 2.5mm thick. To correct
141 for this, `dcm2niix` interpolates between slices to ensure each image has a consistent voxel size. Again,
142 `dcm2niix` returns both the corrected and non-corrected image.

143 Once converted to NIfTI format, one should ensure the scale of the data. Most CT data is between
144 -1024 and 3071 Hounsfield Units (HU). Values less than -1024 HU are commonly found due to areas
145 of the image outside the field of view that were not actually imaged. One first processing step would be
146 to Winsorize the data to the $[-1024, 3071]$ range. After this step, the header elements `scl_slope` and
147 `scl_inter` elements of the NIfTI image should be set to 1 and 0, respectively, to ensure no data rescaling
148 is done in other software. Though HU is the standard format used in CT analysis, negative HU values may
149 cause issues with standard imaging pipelines built for MRI, which typically have positive values. Rorden
150 (CITE) proposed a lossless transformation, called Cormack units, which have a minimum value of 0. The
151 goal of the transformation is to increase the range of the data that is usually of interest, from -100 to
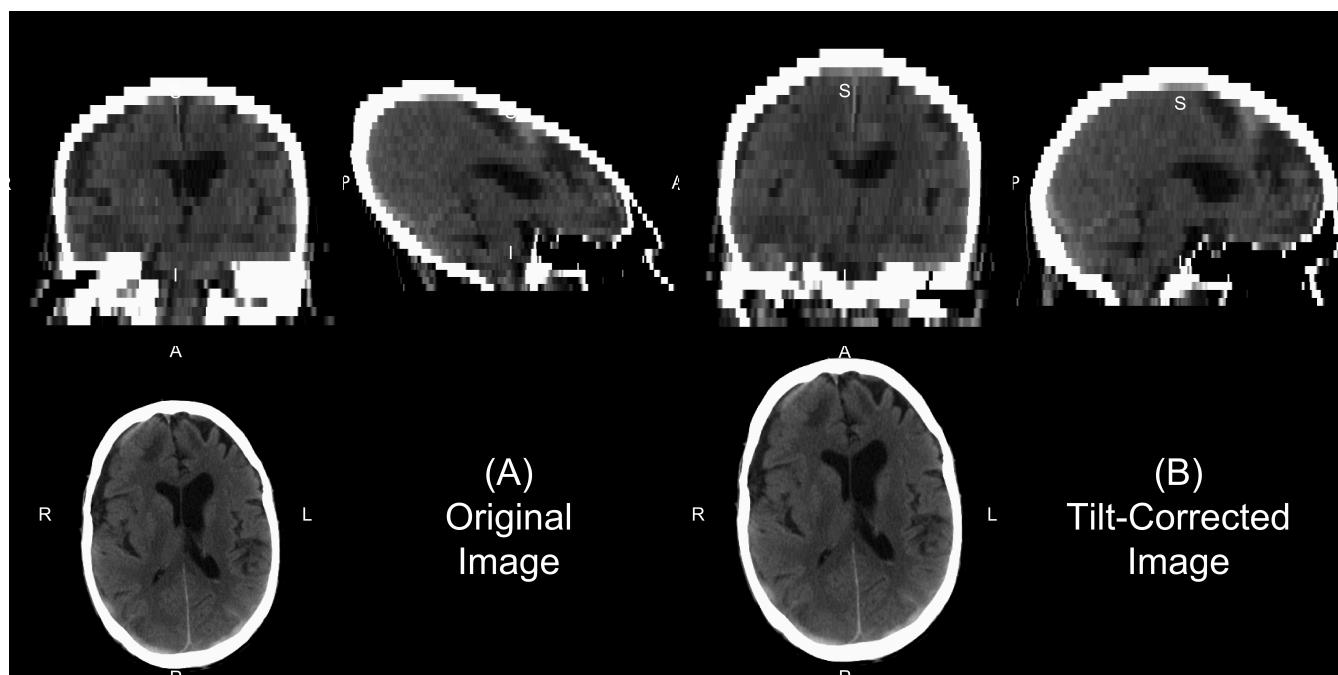
Figure 1. Example of Gantry-tilt Correction. Using 'dcm2niix', we converted the DICOM files to a NIfTI file, which had a 30 degree tilt. The output provides the uncorrected image (A) and the tilt-corrected image (B). We see that the reconstructed image without correction appears fine within the axial plane, but out of plane has an odd 3D shape. This shape will be corrected with an affine transformation, which is done in the conversion, correcting the image as seen in (B).

152   100HU and is implemented in the Clinical toolbox (discussed below). Most analyses are done using HU,
153   however.

2.4   Convolution Kernel

155   Though we discuss CT as having more standardized Hounsfield unit values, this does not imply CT scans
156   cannot have vastly different properties depending on parameters of scanning and reconstruction. One notable
157   parameter in image reconstruction is the convolution kernel (i.e. filter, DICOM field `(0018,1210)`)
158   used for reconstruction. We present slices from an individual subject from the CQ500 (Chilamkurthy et
159   al., 2018) dataset in Figure 2. Information on which kernel was used, and other reconstruction parameter
160   information can be found in the DICOM header. The kernel is described usually by the letter "H" (for
161   head kernel), a number indicating image sharpness (e.g. the higher the number, the sharper the image,
162   the lower the number, the smoother the image), and an ending of "s" (standard), "f" (fast), "h" for high
163   resolution modes (Siemens SOMATOM Definition Application Guide), though some protocols simply
164   name them "soft-tissue", "standard", "bone", "head", or "blood", amongst others. The image contrast can
165   depend highly on the kernel, and "medium smooth" kernels (e.g. H30f, H30s) can provide good contrast
166   in brain tissue (Figure 2E). Others, such as "medium" kernels (e.g. H60f, H60s) provide contrast in high
167   values of the image, such as detecting bone fractures (Figure 2A), but not as good contrast in brain tissue
168   (Figure 2B). Thus, when combining data from multiple sources, the convolution kernel may be used to
169   filter, stratify, or exclude data.

170   Moreover, the noise and image contrast can be different depending on the image resolution of the
171   reconstruction. Most standard head CT scans have high resolution within the axial plane (e.g. 0.5x0.5mm).
172   Image reconstructions can have resolution in the interior-superior plane (e.g. slice thickness) anywhere
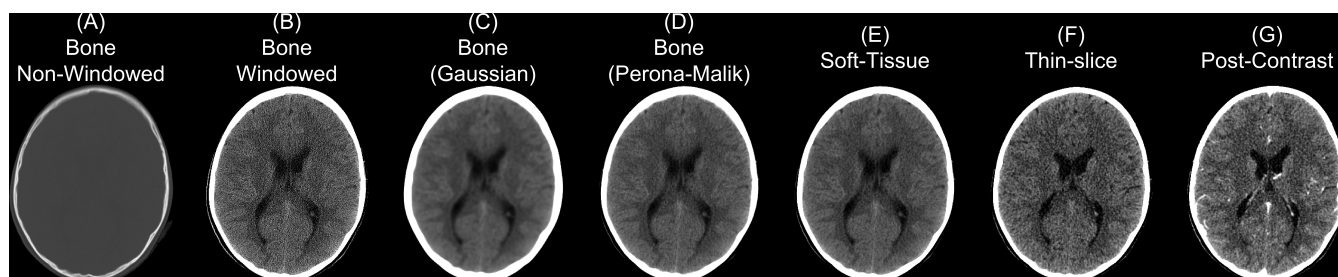
Figure 2. Different Series for a Scanning Study. Here we present different non-contrast head CT exported from a PACS. We display a reconstructed scan with a bone convolution kernel (A), showing bright areas of the skull, which can be investigated for breaks or fractures. When applying a window of 0 - 100 Hounsfield units (HU) of this image, we see the image resolution (B). Using a Gaussian (C) or Perona-Malik (D) smoother, we see we get smoother images, similar to the image reconstructed with a soft-tissue convolution kernel (E). Images (A-E) had a slice thickness of 5mm. The thin-slice scan (F) had a slice thickess of 0.62mm and a soft-tissue convolution kernel.

173 from 0.5mm (aka 'thin-slice', Figure 2F)) to 2.5mm, to 5mm, where 5mm is fairly common. The larger
174 the slice thicknesses are, the smoother the reconstruction (as areas are averaged). Also, the added benefit
175 for radiologists and clinicians are that fewer slices are needed to be reviewed for pathology or to get a
176 comprehensive view of the head. In research, however, these thin-slice scans can get better estimates of
177 volumes of pathology, such as a hemorrhage (CITE), or other brain regions. Moreover, when performing
178 operations across images, algorithms may need to take this differing resolution, and therefore image
179 dimensions, into account. We will discuss image registration in the data preprocessing as one way to
180 harmonize the data dimensions, but registration does not change the inherent smoothness or resolution of
181 the original data.

182 In some instances, only certain images are available for certain subjects. For example, most of the
183 subjects have a non-contrast head CT with a soft-tissue convolution kernel, whereas some only have
184 a bone convolution kernel. Post-processing smoothing can be done, such as 3D Gaussian (Figure 2C)
185 or anisotropic (Perona-Malik) smoothing (Perona and Malik, 1990) (Figure 2D). This process changes
186 the smoothness of the data, contrast of certain areas, can cause artifacts in segmentation, but can make
187 the within-plane properties similar for scans with bone convolution kernel reconstructions compared to
188 soft-tissue kernels in areas of the brain (Figure 2E).

189 **2.5 Contrast Agent**

190 Though we are discussing non-contrast scans, head CT scans with contrast agent are common. The
191 contrast/bolus agent again should be identified in the DICOM header field `(0018,0010)`, but may be
192 omitted. The contrast changes CT images, especially where agent is delivered, notably the vascular system
193 of the brain (Figure 2G). These changes may affect the steps recommended in the next section of data
194 preprocessing, where thresholds may need to be adjusted to include areas with with contrast which can
195 have higher values than the rest of the tissue (e.g. > 100HU). (FIGURE)

3 DATA PREPROCESSING

196 Now that the data is in a standard file format, we can discuss data preprocessing. As the data are in NIfTI
197 format, most software built for MRI and other imaging modalities should work, but adaptations and other
198 considerations may be necessary.

3.1 Bias-field/Inhomogeneity Correction

In MRI, the scan may be contaminated by a bias field or set of inhomogeneities. This field is generally due to inhomogeneities/inconsistencies in the MRI coils or can be generated by non-uniform physical effects on the coils, such as heating. One of the most common processing steps done first is to remove this bias field. In many cases, these differences can more general be considered non-uniformities, in the sense that the same area with the same physical composition and behavior may take on a different value if it were in a different spatial location of the image. Though CT data has no coil or assumed bias field due to the nature of the data, one can test if trying to harmonize the data spatially with one of these correction procedures improves performance of a method. Though we do not recommend this procedure generally, as it may reduce contrasts between areas of interest, such as hemorrhages in the brain, but has been used to improve segmentation (Cauley et al., 2018). We would like to discuss potential methods and CT-specific issues.

Overall, the assumptions of this bias field are that it is multiplicative and is smoothly varying. One of the most popular inhomogeneity corrections are the N3 (Sled et al., 1998) an its updated improvement N4 (Tustison et al., 2010) in ANTs, though other methods exist in FSL (Zhang et al., 2001) and other software (Ashburner and Friston, 1998; Belaroussi et al., 2006). Given the assumption of the multiplicative nature of the field, N4 performs an expectation–maximization (EM) algorithm on the log-transformed image, assuming a noise-free system. As CT data in HU has negative values, the log transform is inappropriate. Pre-transforming or shifting the data values may be necessary to perform this algorithm, though these transforms may affect performance. Moreover, artifacts or objects (described below), such as the bed, may largely effect the estimation of the field and segmentation may be appropriate before running these corrections, such as brain extraction or extracting only subject-related data and not imaged hardware. The `ANTsR` package (`https://github.com/ANTsX/ANTsR`) provides the `n4BiasFieldCorrection` function in `R`; `ANTsPy` (`https://github.com/ANTsX/ANTsPy`) and NiPype (Gorgolewski et al., 2011) provide `n4_bias_field_correction` and `N4BiasFieldCorrection` in `Python`, respectively.

3.2 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/gurney, and any instruments in the field of view. We do not provide a general framework to extract the complete head from hardware, but provide some recommendations for working heuristics. Typically the range of data for the brain and facial tissues are within -100 to 300HU, excluding the skull, other bones, and calcificiations. Creating a mask from values from the -100 to 1000HU range tends to remove some instruments, the pillow, and the background. Retaining the largest connected component will remove high values such as the bed/gurney, filling holes (to include the skull), and masking the original data with this resulting mask will return the subject (Figure 3).

Note, care must be taken whenever a masking procedure is used as one standard way is to set values outside an area of interest to 0. With CT data 0 HU is a real value of interest: if all values are set to 0 outside the mask, the value of 0 is aliased 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 values not of interest.

One of the most common steps in processing imaging of the brain is to remove non-brain structures from the image. Many papers present brain extracted CT images, but do not always disclose the method of extraction. We have published a method that uses the brain extraction tool (BET) from FSL, originally
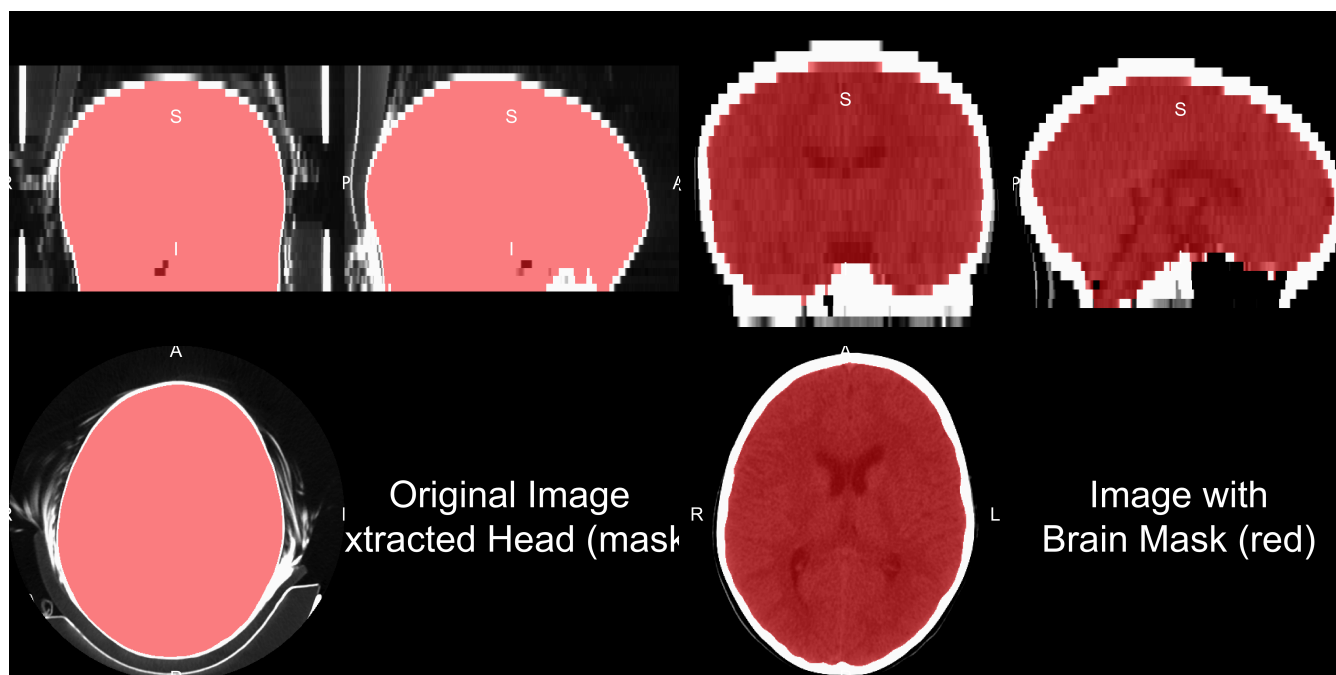
Figure 3. Human and Brain Extraction Results. Here we present a 5mm slice, non-contrast head CT with a soft-tissue convolution kernel. The left figure represents the CT image, showing all the areas imaged, overlaid with the extracted head mask as described in the section of "Brain Extraction in CT". The right hand side is the image overlaid with a brain mask. The brain mask was created using an adaptation of the Brain Extraction Tool (BET) from FSL, published by Muschelli et al. (2015).

built for MRI, to perform brain extraction (Muschelli et al., 2015) with the `CT_Skull_Strip` function in the `ichseg` R package (Muschelli, 2019). An example of this algorithm performance on a 5mm slice, non-contrast head CT with a soft-tissue convolution kernel is seen in Figure 3, which extracts the relevant areas for analysis. 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 traumatic brain injury (TBI) and surgery, such as craniotomies or craniectomies, can cause these methods to potentially fail. Overall, however, large contrast between the skull and brain tissue and standardized Hounsfield Units can make brain segmentation an easier task than in MRI.

3.3 Registration to a CT template

Though many analyses in clinical data may be subject-specific, population-level analyses are still of interest. Some analyses want spatial results at the population-level, which require registration to a population template. One issue with these approaches is that most templates and approaches rely on an MRI template. These templates were developed by taking MRI scans of volunteers, which again is likely unethical with CT due to the radiation exposure risk without other benefits. To create templates, retrospective searches through medical records can provide patients who came in with symptoms warranting a CT scan, such as a migraine, 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
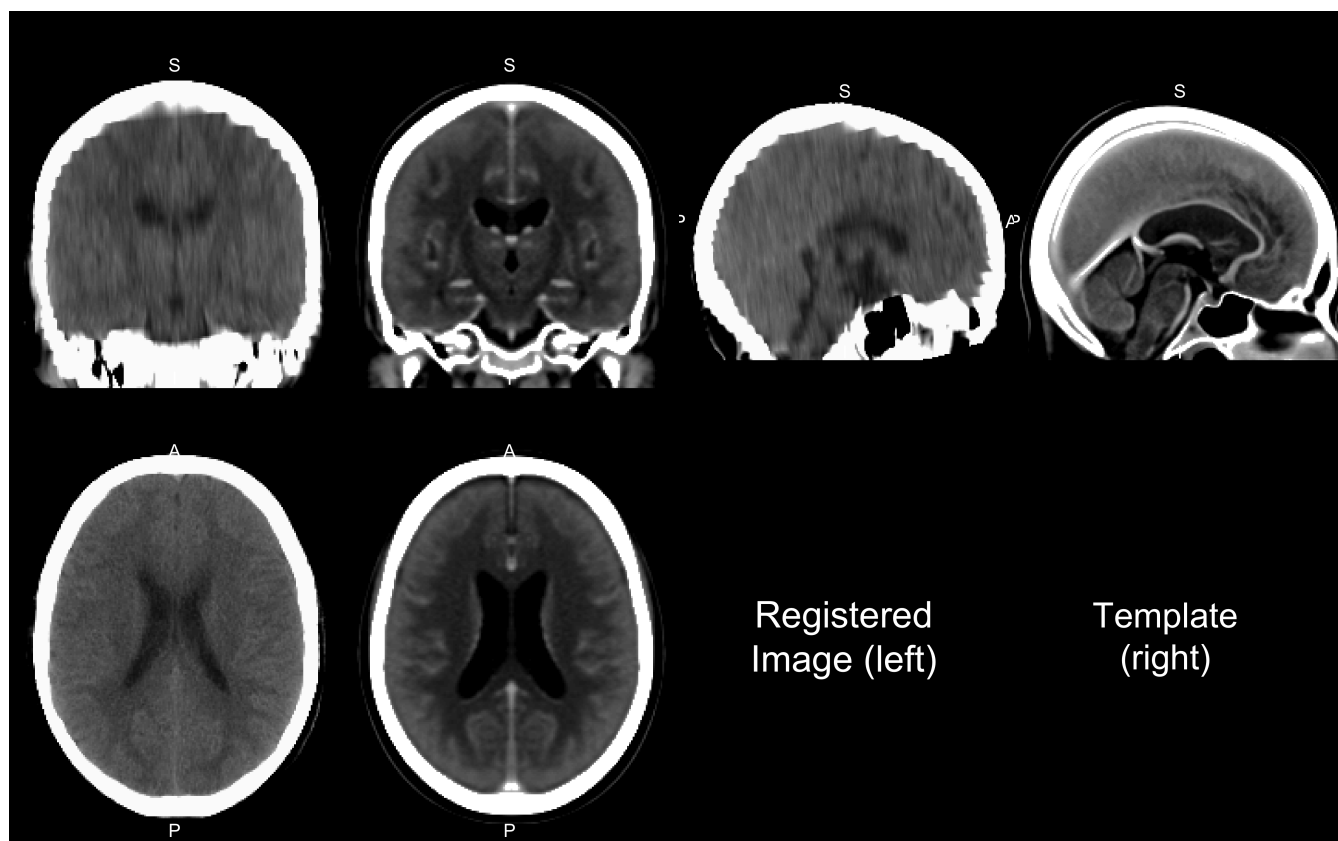
Figure 4. Image Registration Result. Here we use the same scan that we performed brain extraction before, and register it to a CT template (Rorden, 2012). We registered the image using symmetric normalization (SyN), a non-linear registration done after affine registration. We see areas of the image that align generally well, but may not be perfect.

263 population and a research consent at an institution, these scans can be used to create templates and atlases
264 based on CT. To our knowledge, the first publicly available head CT template exists was released by Rorden
265 et al. (2012), for the purpose of spatial normalization/registration.

266   One interesting aspect of CT image registration is again that CT data has units within the same range. To
267 say they are uniformly standardized is a bit too strong as tomography and other confounds can impact units.
268 Thus, it is our practice to think of them as more standardized than MRI. This standardization may warrant
269 or allow the user different search and evaluation cost functions for registration, such as least squares. We
270 have found that normalized mutual information (NMI) still performs well in CT-to-CT registration and
271 should be at least considered when using CT-to-MRI or CT-to-PET registration. Along with the template
272 above, Rorden et al. (2012) released the Clinical toolbox (`https://github.com/neurolabusc/`
273 `Clinical`) for SPM to allow researchers to register head CT data to a standard space. However, as
274 the data are in the NIfTI format, almost all image registration software should work, though one should
275 consider transforming the units using Cormack units or other transformations as negative values may
276 implicitly be excluded in some software built for MRI registration. We have found using diffeomorphic
277 registrations such as symmetric normalization (SyN) from `ANTs` and `ANTsR` with NMI cost functions to
278 perform well. We present results of registering the head CT presented in brain extraction to the template
279 from Rorden et al. (2012) using SyN in Figure 4.

280 In some cases, population-level analyses can be done, but while keeping information at a subject-specific
281 level. For example, registration from a template to a subject space can provide information about brain
282 structures that can be aggregated across people. For example, one can perform a label fusion approach
283 to CT data to infer the size of the hippocampus and then analyze hippocampi sizes across the population.
284 Numerous label fusion approaches exist (Wang et al., 2013; Sabuncu et al., 2010; Asman and Landman,
285 2013; Langerak et al., 2010; Collins and Pruessner, 2010), but rely on multiple templates and publicly
286 available segmented CT images are still lacking. Additionally, the spatial contrast in CT is much lower than
287 T1-weighted MRI for image segmentation. Therefore, concurrent MRI can be useful. One large issue is
288 that any data gathered with concurrent MRI the high variability in MRI protocol done if it is not generally
289 standardized within or across institution. We see these limits as a large area of growth and opportunity in
290 CT image analysis.

291 ### 3.3.1 Concurrent MRI

292 ## 3.4 Pipeline

293 Overall, our recommended pipeline is as follows:

294 1. Use CTP or `DicomCleaner` to organize and anonymize the DICOM data from a PACS.
295 2. Extract relevant header information for each DICOM, using software such as `dcmdump` from `dcmtk`
296   and store, excluding PHI.
297 3. Convert DICOM to NIfTI using `dcm2niix`, which can create brain imaging data structure (BIDS)
298   formatted data (Gorgolewski et al.). Use the tilt-corrected and data with uniform voxel size.

299 After, depending on the purpose of the analysis, you may do registration then brain extraction, brain
300 extraction then registration, or not do registration at all. If you are doing analysis of the skull, you can also
301 use brain extraction as a first step to identify areas to be removed. For brain extraction, run `BET` for CT or
302 `CT_BET` (especially if you have GPUs for the neural network). If registration is performed, keeping the
303 transformations back into the native, subject space is usually necessary as many radiologists and clinicians
304 are comfortable to subject-specific predictions or segmentations. Converting the data from NIfTI back to
305 DICOM is not commonly done, but is possible as most PACS are built for DICOM data.

4 CONCLUSIONS

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

314 For research using head CT scans to have the level of interest and success as MRI, additional publicly
315 available data needs to be released. We saw the explosion of research in MRI, particularly functional MRI,
316 as additional data were released and and consortia created truly large-scale studies. This collaboration is
317 possible at an individual institution, but requires scans to be released from a clinical population, where
318 consent must be first obtained, and upholding patient privacy must be a top priority. Large internal data sets
319 likely exist, but institutions need incentives to release these data sets to the public. Also, though institutions

320 have large amounts of rich data, general methods and applications require data from multiple institutions
321 as parameters, protocols, and population characteristics can vary widely.

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

327   All of the code used to generate the figures in this paper is located at `https://github.com/`
328 `muschellij2/process_head_ct`. The code uses packages from Neuroconductor in `R`. All data
329 presented was from the CQ500 data set, which can be downloaded from `http://headctstudy.qure.`
330 `ai/dataset`.

REFERENCES

331 Akkus, Z., Kostandy, P. M., Philbrick, K. A., and Erickson, B. J. (2018). Extraction of brain tissue from
332 CT head images using fully convolutional neural networks. in *Medical imaging 2018: Image processing*
333 (International Society for Optics; Photonics), 1057420.

334   Aryanto, K., Oudkerk, M., and Ooijen, P. van (2015). Free DICOM de-identification tools in clinical
335 research: Functioning and safety of patient privacy. *European radiology* 25, 3685–3695.

336   Ashburner, J., and Friston, K. (1998). MRI sensitivity correction and tissue classification. *NeuroImage* 7.

337   Asman, A. J., and Landman, B. A. (2013). Non-local statistical label fusion for multi-atlas segmentation.
338 *Medical image analysis* 17, 194–208.

339   Belaroussi, B., Milles, J., Carme, S., Zhu, Y. M., and Benoit-Cattin, H. (2006). Intensity non-uniformity
340 correction in MRI: Existing methods and their validation. *Medical image analysis* 10, 234–246.

341   Cauley, K. A., Och, J., Yorks, P. J., and Fielden, S. W. (2018). Automated segmentation of head computed
342 tomography images using FSL. *Journal of computer assisted tomography* 42, 104–110.

343   Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., Mahajan, V.,
344 Rao, P., and Warier, P. (2018). Deep learning algorithms for detection of critical findings in head ct scans:
345 A retrospective study. *The Lancet* 392, 2388–2396.

346   Clayden, J., and Rorden, C. (2018). *divest: Get images out of DICOM format quickly*. Available at:
347 `https://CRAN.R-project.org/package=divest`.

348   Collins, D. L., and Pruessner, J. C. (2010). Towards accurate, automatic segmentation of the hippocampus
349 and amygdala from mri by augmenting ANIMAL with a template library and label fusion. *Neuroimage* 52,
350 1355–1366.

351   Eichelberg, M., Riesmeier, J., Wilkens, T., Hewett, A. J., Barth, A., and Jensch, P. (2004). Ten years of
352 medical imaging standardization and prototypical implementation: The DICOM standard and the OFFIS
353 DICOM toolkit (DCMTK). in *Medical imaging 2004: PACS and imaging informatics* (International Society
354 for Optics; Photonics), 57–69.

355   Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh,
356 S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in
357 Python. *Frontiers in neuroinformatics* 5, 13.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data* 3, 160044 EP. Available at: `https://doi.org/10.1038/sdata.2016.44`.

Langerak, T. R., Heide, U. A. van der, Kotte, A. N., Viergever, M. A., Van Vulpen, M., Pluim, J. P., and others (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE transactions on medical imaging* 29, 2000–2008.

Li, X., Morgan, P. S., Ashburner, J., Smith, J., and Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of neuroscience methods* 264, 47–56.

Mason, D. (2011). SU-e-t-33: pydicom: An open source DICOM library. *Medical Physics* 38, 3493–3493.

Muschelli, J. (2018). *dcm2niir: Conversion of DICOM to NIfTI imaging files through R*. Available at: `https://github.com/muschellij2/dcm2niir`.

Muschelli, J. (2019). *ichseg: Intracerebral hemorrhage segmentation of x-ray computed tomography (CT) images*.

Muschelli, J., Gherman, A., Fortin, J. P., Avants, B., Whitcher, B., Clayden, J. D., Caffo, B. S., and Crainiceanu, C. M. (2018). Neuroconductor: An R platform for medical imaging analysis. *Biostatistics*.

Muschelli, J., Ullman, N. L., Mould, W. A., Vespa, P., Hanley, D. F., and Crainiceanu, C. M. (2015). Validated automatic brain extraction of head CT images. *Neuroimage* 114, 379–385.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier.

Perona, P., and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence* 12, 629–639.

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: `https://www.R-project.org/`.

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., and Karnath, H.-O. (2012). Age-specific CT and MRI templates for spatial normalization. *Neuroimage* 61, 957–965.

Russell, P. (2018). *TCIApathfinder: Client for the cancer imaging archive rest api*. Available at: `https://CRAN.R-project.org/package=TCIApathfinder`.

Russell, P. H., and Ghosh, D. (2019). radtools: R utilities for convenient extraction of medical image metadata. *F1000Research* 7.

Sabuncu, M. R., Yeo, B. T., Van Leemput, K., Fischl, B., and Golland, P. (2010). A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging* 29, 1714–1729.

Schroeder, W., Ng, L., and Cates, J. (2003). The ITK software guide.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging* 17, 87–97.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE transactions on medical imaging* 29, 1310.

395   Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2013). Multi-atlas
396   segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence* 35,
397   611–623.

398   Whitcher, B., Schmid, V. J., and Thornton, A. (2011). Working with the DICOM and NIfTI data standards
399   in R. *Journal of Statistical Software*.

400   Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov
401   random field model and the expectation-maximization algorithm. *Medical Imaging, IEEE Transactions on*
402   20, 45–57. Available at: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=`
403   `906424` [Accessed September 9, 2014].