International Journal of Clinical Trials
Harris RJ et al. Int J Clin Trials. 2018 Nov;5(4):161-169
http://www.ijclinicaltrials.com

pISSN 2349-3240 | eISSN 2349-3259

DOI: http://dx.doi.org/10.18203/2349-3259.ijct20184398

Original Research Article

High-throughput image labeling and quality control for clinical trials using machine learning

Robert J. Harris[1], Pangyu Teng[2], Mahesh Nagarajan[2], Liza Shrestha[2], Xiang Lu[1], Bharath Ramakrishna[1], Peiyun Lu[1], Theo Sanford[1], Heather Clem[1], Megan McRoberts[1], Jonathan Goldin[1], Matt Brown[1]*

[1]MedQIA LLC, Los Angeles, CA, USA
[2]Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Received: 16 May 2018
Accepted: 05 October 2018

***Correspondence:**
Dr. Matt Brown,
E-mail: mbrown@medqia.com

ABSTRACT

Background: Manually importing and analyzing image data can be time-consuming, prone to human error, and costly for large clinical trial datasets. This can lead to delays in quality control (QC) feedback to imaging sites and in obtaining data analysis results. Herein we describe the creation and application of a high-throughput review process for import, classification, labeling and QC of large multimodal clinical trial image datasets.
Methods: Automated methods were used to remove patient identifying information, extract image header data, and filter image data for usability. A convolutional neural net was applied to estimate anatomy for CT images. Internal scores were assigned for each image series to identify the optimal series for labeling and reading of each anatomical region. Image QC reports were automatically generated for all patients.
Results: In combined studies for which 204,492 series were received, 27,841 series were identified as usable and 13,415 series were labeled. Using this high-throughput method, total work-hours required per time point were reduced by an approximate factor of ten when compared to traditional review and labeling methods. Our anatomic classification system identified 95.7% of image series correctly, with the remaining series being manually corrected before labeling and analysis.
Conclusions: A high-throughput image analysis pipeline was implemented in a large combined dataset of clinical trial image series. This pipeline can be applied across other studies and modalities for fast image data characterization, labeling and QC.

Keywords: Image intake, High-throughput, Machine learning, DICOM, Data management

INTRODUCTION

Clinical trial data collected from multiple imaging sites can often be heterogeneous, redundant, or sometimes unusable. To detect the underlying treatment signal in a patient cohort, the data must first be organized as cleanly and homogenously as possible before starting analysis.[1] For prior studies undertaken by our organization, a standard workflow was to initially import received image data into our in-house image storage and viewer system database. Once the data is present in the database, typically each imaging study time point is manually reviewed to identify the modality, anatomy, and other characteristics and select the best image series (or set of series) for labeling. This requires opening and reviewing each image series, which is a time consuming process given that each time point will typically include multiple image series including scout views, multiple reconstructions, and reformats. From among all of those series only those that need to be read are labeled.

Therefore, each time point has typically taken approximately 30 minutes to check for and remove protected health information (PHI), then import, review, and label.

PHI cleaning can be particularly time consuming, as PHI can reside in both known and private DICOM tags or be burnt into the image data itself. For a given reading task, the optimal scans from those available at each time point based on both anatomical coverage and acquisition parameters must be selected and labeled for reading. Using a high-throughput technique for characterizing and labeling data is thereby warranted, especially for large studies where manual review of the high number of image series is not feasible, such as retrospective studies where a large bolus of imaging data may arrive at once.

A number of scientific fields require high-throughput analysis of large datasets, especially in the biological field.[2–4] As such, the use of computational data management solutions continues to be explored in a variety of settings to improve the efficiency and accuracy of data intake and analysis.[5,6] While these methods often apply to datasets involving millions of data points, the same principles can be applied to image datasets made up of hundreds of thousands of image series. High-throughput pipelines for processing biologic image data typically employ machine learning and data mining techniques for image classification and analysis.[7] Medical image data typically arrives in DICOM format; it comprises both the image data and a header file containing scan parameters and other identifying image data. Assigning labels to organize image data can be posed as a classification problem that uses both the image and header information as input while also applying machine learning.

Herein we describe the development and application of a novel, high-throughput pipeline for quality control (QC) and labeling of medical images in clinical trials. In this context, labeling refers to marking an image series in the database by its modality, anatomy, and other characteristics such as the presence of gadolinium contrast agent; the required types of labeled series are often defined by the study or clinical trial protocol and imaging charter. This pipeline can be used to screen large amounts of image data to identify and label only the optimal image series for each modality and anatomy. These optimal labeled scans can then be passed to radiologists for analysis. We hypothesized that an automated high-throughput pipeline could accurately classify imaged anatomy, assign QC scores to each series, tag images for labeling and analysis, and calculate an overall quality score for each submitted time point, with few labeling corrections being required by a human image analyst.

METHODS

The data present in this study consisted largely of computed tomography (CT) and nuclear medicine (NM) technetium-99m (Tc-99m) bone scan scintigraphy images pooled from three clinical trial datasets: one prostate cancer therapy trial and two lymphoma therapy trials. Although the NM category of imaging can encompass a range of techniques including position emission tomography (PET) and single-photon emission computed tomography (SPECT) imaging, "NM" will refer to Tc-99m bone scan acquisitions throughout this study. CT data is typically multi-slice and can be reconstructed in the axial, sagittal, or coronal planes. For the purposes of this study, only axial CT series were considered usable, as this is the standard image orientation used by radiologists for quantitative analysis. NM data may arrive as whole-body bone scan images, which were the ideal form of NM data for this study; screen captures, which are a secondary form of NM data that may combine multiple whole-body bone scans; or a number of spot views and 3D reconstructions, which were not usable for this study. As usable whole-body bone scan data is inherently 2D rather than 3D, ideal NM data should have only one DICOM image per series. Therefore, it was first necessary to split any NM series containing multiple DICOM images into an individual series per DICOM image prior to import. This splitting of NM series occurred as data was entering the import pipeline. The goal of this data QC and labeling pipeline was to select the optimal CT and NM series for labeling and prepare them for response evaluation criteria in solid tumors (RECIST) or prostate cancer working group (PCWG2) evaluation, respectively, which are standardized criteria for evaluating data from those modalities.[8,9]

Our approach to high-throughput image intake and labeling combines: (1) automated anatomic and technical parameter classification, and (2) human review processes and HTML user interfaces. The pipeline workflow is shown in Figure 1 and subsequently described in detail.

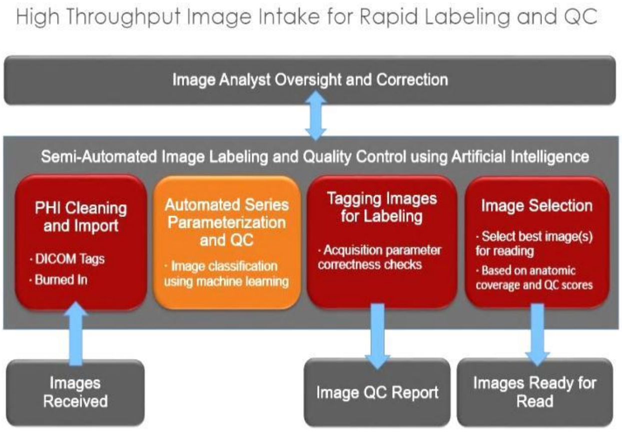
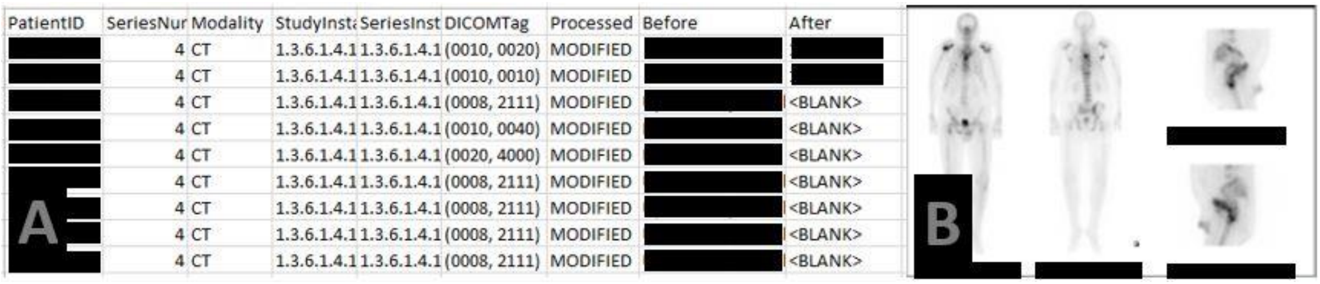

High Throughput Image Intake for Rapid Labeling and QC
Image Analyst Oversight and Correction
Semi-Automated Image Labeling and Quality Control using Artificial Intelligence
PHI Cleaning and Import
· DICOM Tags
· Burned In
Automated Series Parameterization and QC
· Image classification using machine learning
Tagging Images for Labeling
· Acquisition parameter correctness checks
Image Selection
· Select best image(s) for reading
· Based on anatomic coverage and QC scores
Images Received
Image QC Report
Images Ready for Read

Figure 1: A flowchart describing the high-throughput image intake and labeling process.

PHI cleaning and import

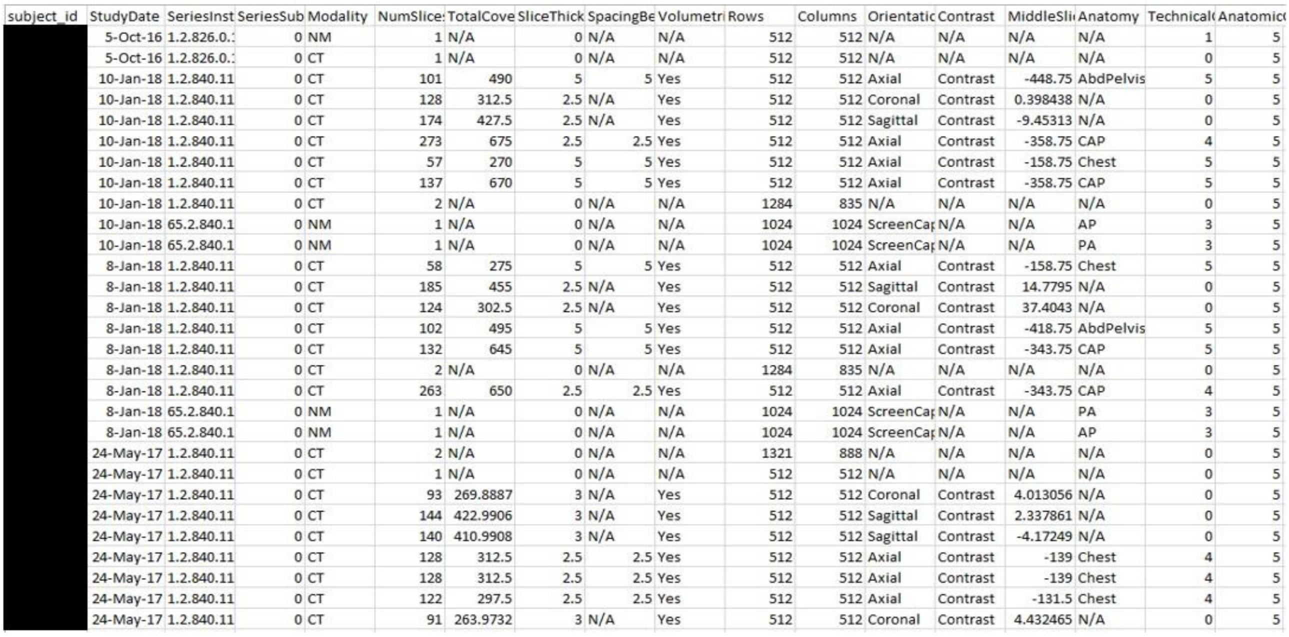
PHI must be removed from images prior to import into the information system. Although trials typically require sites to remove PHI before transfer to the central radiology facility, it can sometimes still be found in DICOM header tags or burnt into the image pixels. Our high-throughput system automatically scrubs the contents of standard tags where PHI can occur, such as the date of birth, patient weight, and patient address tags, as well as private tags that are freely used by scanner manufacturers. The list of DICOM tags to be scrubbed is contained in a configuration file that can be adjusted depending on study requirements. The system then generates a log file showing all fields that were scrubbed (Figure 2A) so that a human image analyst can review and report any PHI that was removed to the site and trial sponsor. This log file contains a row for each series that contained DICOM header PHI, with each row providing the subject ID, modality, study instance UID, series instance UID, DICOM tag number, and the PHI within that DICOM tag. This log file is then manually reviewed to determine whether the identified DICOM tags actually contain PHI and verify the type of PHI. A "Type of PHI" field is manually entered into the log file containing the PHI category (accession number, subject ID, etc.). A "type of PHI Location" field is also manually entered into the log file, which differs from the Type of PHI field only in cases where a mismatch occurs between the DICOM tag and the type of PHI (e.g. an address in the patient name DICOM tag).

Burnt-in PHI is primarily present in reports saved as DICOM files or reformatted images, i.e. series having a small number of image slices. Therefore the system generated scrollable HTML pages of series having fewer than eight slices. The cutoff of eight slices was empirically determined based on previous observations of image series that did or did not contain PHI; as series containing greater than eight slices almost always contain true image data rather than these reports or reformatted images, these series are unlikely to ever contain burnt-in PHI. These HTML pages allowed rapid manual review for PHI that could be and redacted as shown (Figure 2B). Following burnt-in PHI redaction, the image data was imported into our information system.



PatientID	SeriesNur	Modality	StudyInst	SeriesInst	DICOMTag	Processed	Before	After
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0010, 0020)	MODIFIED	[redacted]	[redacted]
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0010, 0010)	MODIFIED	[redacted]	[redacted]
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0008, 2111)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0010, 0040)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0020, 4000)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0008, 2111)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0008, 2111)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0008, 2111)	MODIFIED	[redacted]	<BLANK>
[redacted]	4	CT	1.3.6.1.4.1	1.3.6.1.4.1	(0008, 2111)	MODIFIED	[redacted]	<BLANK>

Figure 2: (A) An example log file generated from DICOM header data to review for PHI; (B) an example HTML thumbnail with PHI text present. Study-specific identifiers have been redacted for this figure (black bars).



subject_id	StudyDate	SeriesInst	SeriesSub	Modality	NumSlice	TotalCove	SliceThick	SpacingBe	Volumetri	Rows	Columns	Orientatic	Contrast	MiddleSli	Anatomy	Technical	Anatomic
[redacted]	5-Oct-16	1.2.826.0.:	0	NM	1	N/A	0	N/A	N/A	512	512	N/A	N/A	N/A	N/A	1	5
[redacted]	5-Oct-16	1.2.826.0.:	0	CT	1	N/A	0	N/A	N/A	512	512	N/A	N/A	N/A	N/A	0	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	101	490	5	5	Yes	512	512	Axial	Contrast	-448.75	AbdPelvis	5	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	128	312.5	2.5	N/A	Yes	512	512	Coronal	Contrast	0.398438	N/A	0	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	174	427.5	2.5	N/A	Yes	512	512	Sagittal	Contrast	-9.45313	N/A	0	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	273	675	2.5	2.5	Yes	512	512	Axial	Contrast	-358.75	CAP	4	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	57	270	5	5	Yes	512	512	Axial	Contrast	-158.75	Chest	5	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	137	670	5	5	Yes	512	512	Axial	Contrast	-358.75	CAP	5	5
[redacted]	10-Jan-18	1.2.840.11	0	CT	2	N/A	0	N/A	N/A	1284	835	N/A	N/A	N/A	N/A	0	5
[redacted]	10-Jan-18	65.2.840.1	0	NM	1	N/A	0	N/A	N/A	1024	1024	ScreenCap	N/A	N/A	AP	3	5
[redacted]	10-Jan-18	65.2.840.1	0	NM	1	N/A	0	N/A	N/A	1024	1024	ScreenCap	N/A	N/A	PA	3	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	58	275	5	5	Yes	512	512	Axial	Contrast	-158.75	Chest	5	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	185	455	2.5	N/A	Yes	512	512	Sagittal	Contrast	14.7795	N/A	0	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	124	302.5	2.5	N/A	Yes	512	512	Coronal	Contrast	37.4043	N/A	0	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	102	495	5	5	Yes	512	512	Axial	Contrast	-418.75	AbdPelvis	5	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	132	645	5	5	Yes	512	512	Axial	Contrast	-343.75	CAP	5	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	2	N/A	0	N/A	N/A	1284	835	N/A	N/A	N/A	N/A	0	5
[redacted]	8-Jan-18	1.2.840.11	0	CT	263	650	2.5	2.5	Yes	512	512	Axial	Contrast	-343.75	CAP	4	5
[redacted]	8-Jan-18	65.2.840.1	0	NM	1	N/A	0	N/A	N/A	1024	1024	ScreenCap	N/A	N/A	PA	3	5
[redacted]	8-Jan-18	65.2.840.1	0	NM	1	N/A	0	N/A	N/A	1024	1024	ScreenCap	N/A	N/A	AP	3	5
[redacted]	24-May-17	1.2.840.11	0	CT	2	N/A	0	N/A	N/A	1321	888	N/A	N/A	N/A	N/A	0	5
[redacted]	24-May-17	1.2.840.11	0	CT	1	N/A	0	N/A	N/A	512	512	N/A	N/A	N/A	N/A	0	5
[redacted]	24-May-17	1.2.840.11	0	CT	93	269.8887	3	N/A	Yes	512	512	Coronal	Contrast	4.013056	N/A	0	5
[redacted]	24-May-17	1.2.840.11	0	CT	144	422.9906	3	N/A	Yes	512	512	Sagittal	Contrast	2.337861	N/A	0	5
[redacted]	24-May-17	1.2.840.11	0	CT	140	410.9908	3	N/A	Yes	512	512	Sagittal	Contrast	-4.17249	N/A	0	5
[redacted]	24-May-17	1.2.840.11	0	CT	128	312.5	2.5	2.5	Yes	512	512	Axial	Contrast	-139	Chest	4	5
[redacted]	24-May-17	1.2.840.11	0	CT	128	312.5	2.5	2.5	Yes	512	512	Axial	Contrast	-139	Chest	4	5
[redacted]	24-May-17	1.2.840.11	0	CT	122	297.5	2.5	2.5	Yes	512	512	Axial	Contrast	-131.5	Chest	4	5
[redacted]	24-May-17	1.2.840.11	0	CT	91	263.9732	3	N/A	Yes	512	512	Coronal	Contrast	4.432465	N/A	0	5

Figure 3: A CSV file is generated from the image data, with each row representing one series in the dataset. Parameters such as slice thickness, orientation, contrast, and anatomy are used to assign a technical QC score to each series. The best series for each time point are then selected for labeling in the label column. Study-specific identifiers such as subject ID have been redacted for this figure (black bars).

Automated anatomic labeling and QC

The automated system automatically computed anatomic labels using a machine learning technique and quality control scores based on image acquisition parameters from the DICOM image header. It outputted a file containing technical acquisition parameters, anatomic coverage information, and a series QC score (Figure 3). Each series parameter calculation is described below. The technical QC score ranging from 0-5 is calculated based on DICOM header values, with 0-2 representing non-usable series (non-axial scans, non-contiguous scans, slice thicknesses above 5 mm, non-usable modalities) and 3-5 representing usable series.

Subject ID: Extracted from DICOM header tag (0010,0020).

Study date: Extracted from DICOM header tag (0008,0020) and modified to DD–MMM–YY format.

Series instance UID: Extracted from DICOM header tag (0020,000e).

Modality: Extracted from DICOM header tag (0008,0060).

Number of slices: Calculated by reading through all series image slices using a *for* loop and iterating a slice counter.

Total coverage (mm): Calculated by subtracting the second slice location from the second-to-last in the series. (Series may have a discontinuous overview image as their first or last slice, so these slices were selected instead.)

Slice thickness (mm) – Taken from DICOM header tag (0018,0050).

Spacing between slices (mm) – Taken from DICOM header tag (0018,0088).

Contiguity: An *Empirical Slice Thickness* was first calculated by subtracting the second slice location in the series from the third location. If this value was less than or equal to Slice Thickness, the series was contiguous, otherwise it was non-contiguous.

Rows: Taken from DICOM header tag (0028,0010).

Columns: Taken from DICOM header tag (0028,0011).

Orientation: For CT series, the array in DICOM header tag (0020,0037) was taken. If we define this as *Array* and its first value is Array[0], orientation was determined by the following rules:

Array [0] >0.90 and Array[4] >0.90, Axial;
Array [1] >0.90 and Array[5] <-0.90, Sagittal;
Array [0] >0.90 and Array[5] <-0.90, Coronal.

For NM series, this orientation column denotes whether the image was an original whole-body scan or a secondary screen capture. The following criteria were empirically determined after observing several test cases. If:

Rows = 1024 and Columns = 256 or 512, the orientation is "Original";
Rows >512 and Columns >512, the orientation is "Screen Capture";
Rows ≤512 or Columns ≤512, the orientation is "N/A".

Contrast: This field was set to "Contrast" by default; during manual review it was changed to "Non-Contrast" where necessary.

Anatomy: For CT series, we determined the anatomy using a deep learning approach involving a convolutional neural net (*keras; theano as backend*). For initial training, we manually categorized axial CT images into one of eight categories: head, shoulder, upper chest, middle chest, lower chest, abdomen, pelvis, and thigh, with each category containing 3300 images. The dataset was split into a training-validation-test set using a ratio of 0.64:0.16:0.20. Image intensity was scaled from [-1000HU,1000HU] to [0.0,1.0] (HU=Hounsfield Units) and the images were down sampled to a resolution of 256x256. The neural network consisted of four convolutional layers and three fully connected layers, similar to that implemented by Roth et al.[10] Max pooling and dropout (probability of 0.5) were used for each layer, while batch normalization was used for all convolutional and fully connected layers. [11-13] A batch size of 32 was used, and ReLU was uniformly used as the activation type.[14] During training, the optimizer used was Adam (learning rate =0.001).[15] Data augmentation included rotating the images and flipping them in both up/down and left/right directions. Training accuracy was 95.2% for the training set, 95.6% for the validation set, and 95.8% for the test set. When implementing the neural net in the high-throughput pipeline, a resulting anatomical classification was assigned to each image series slice. For each type of anatomy, the number of slices categorized into that anatomy was then multiplied by slice thickness to obtain an anatomy length (mm) for that series. Based on empirical observations of a data subset, rules were developed to automatically label the CT series into one of the following categories in Table 1. The table cells show the length of the scan classified as a particular anatomic region (Upper Chest, Abdomen, Pelvis) by the neural net and the corresponding label assigned by the system.

For NM series, if the DICOM header tag (0054,0400) started with an A (e.g. AP, Ant, Anterior), the anatomy was classified as Anterior–Posterior (AP). A Posterior–Anterior (PA) classification was given when the tag began with P (e.g. PA, Post, Posterior). If the tag was not populated, it was classified as Unknown, with manual review used to determine the anatomy.

Table 1: Anatomic labeling rules for CT series anatomy based on neural net output.

Upper chest (mm)	Abdomen (mm)	Pelvis (mm)	Anatomic label
≥50	≥100	≥100	Chest–Abdomen–Pelvis
≥50	≥140	<100	Chest–Abdomen
<50	≥100	≥100	Abdomen–Pelvis
≥50	<140	<100	Chest
<50	≥100	<100	Abdomen
<50	<100	≥100	Pelvis
<50	<100	<100	Other

Table 2: The effect of slice thickness on the technical QC Score of CT images.

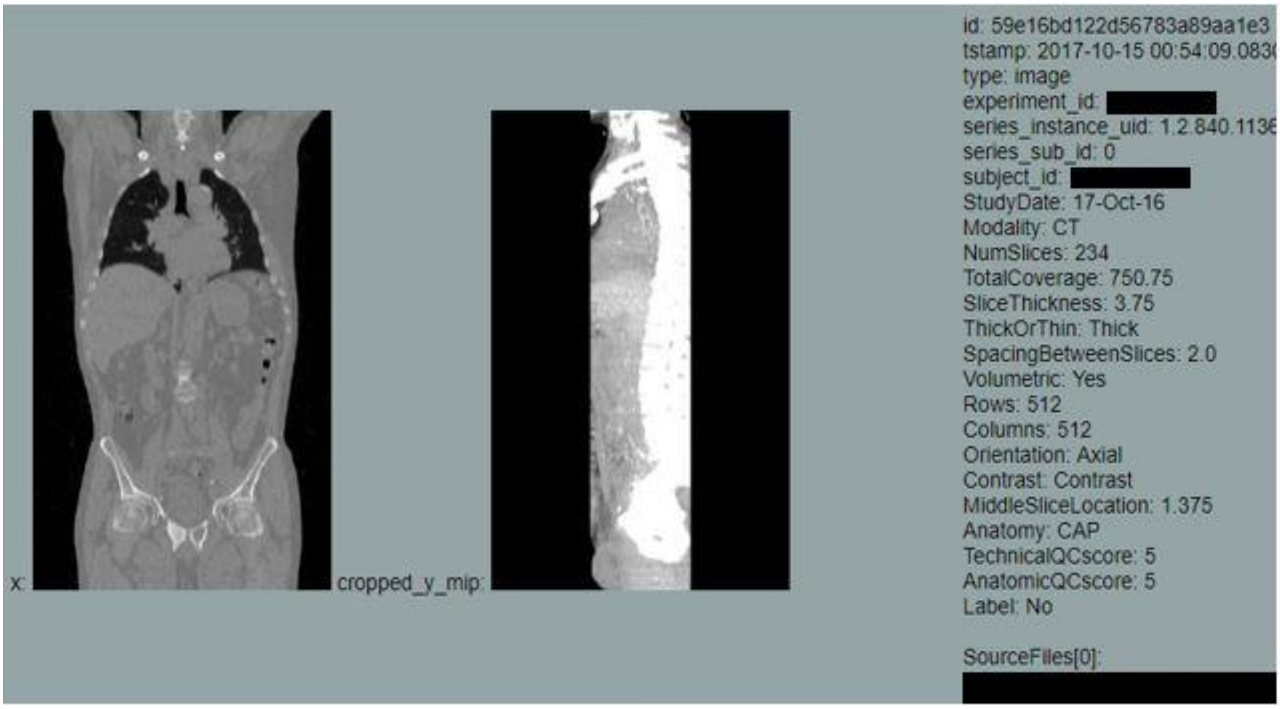
Slice Thickness (mm)	≤3	>3 and ≤5	>5 and ≤10	>10
Slice thickness QC score component	4	5	3	1

*Technical QC score***:** Each series was assigned a technical QC score ranging from 0 to 5, with 5 being the highest score. For CT series, this score was based on score components derived from the slice thickness, contrast, contiguity, and orientation fields. The slice thickness QC score criteria are shown in Table 2, with slice thicknesses between 3 mm and 5 mm defined as the optimal balance between slice resolution and image signal-to-noise for radiologists.

If a CT series was non-contrast, the contrast QC score component was 3, or was 5 otherwise. The contiguity QC score component was 0 if the series was non-contiguous or contiguity could not be determined, or was 5 otherwise. The orientation QC score component was 5 if the series was axial, or was 0 otherwise. The overall technical QC score for CT was then the minimum of these four categories. For NM series, the technical QC score was determined by whether it was an original image or screen capture. If the series was an original image, the score was 5; if the series was a screen capture, the score was 3; otherwise, the score was 0.

Manual review and editing of the classification results

Next, the automated labeling and QC results were used to create an HTML page containing information for each series (Figure 4). During this step, a sagittal maximum intensity projection (MIP) image was calculated for all usable CT series to visually assess whether contrast was present in the tissue.[16] The HTML pages were filtered so as to display only CT series having an axial orientation, and only NM series having a computed technical QC Score ≥3, as series with technical QC scores below 3 were not candidates for labeling.



id: 59e16bd122d56783a89aa1e3
tstamp: 2017-10-15 00:54:09.083
type: image
experiment_id: [redacted]
series_instance_uid: 1.2.840.113
series_sub_id: 0
subject_id: [redacted]
StudyDate: 17-Oct-16
Modality: CT
NumSlices: 234
TotalCoverage: 750.75
SliceThickness: 3.75
ThickOrThin: Thick
SpacingBetweenSlices: 2.0
Volumetric: Yes
Rows: 512
Columns: 512
Orientation: Axial
Contrast: Contrast
MiddleSliceLocation: 1.375
Anatomy: CAP
TechnicalQCscore: 5
AnatomicQCscore: 5
Label: No

SourceFiles[0]:
[redacted]

Figure 4: A thumbnail, MIP, and set of image series data for a chest-abdomen-pelvis series. The neural net classifier output is shown in the anatomy field. Study-specific identifiers have been redacted for this figure (black bars).

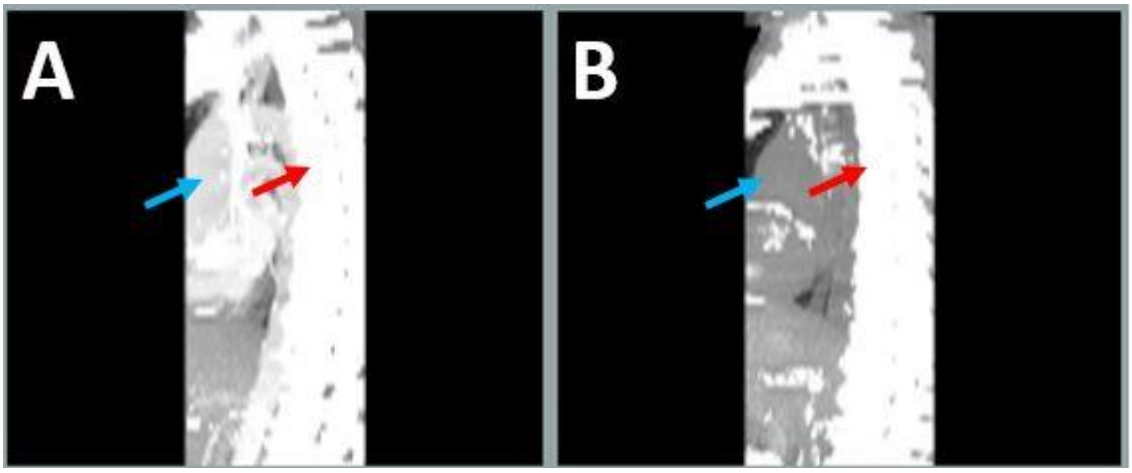


Figure 5: MIP images on the HTML page showing: (A) post-contrast chest CT series and (B) non-contrast chest CT series. A lack of contrast enhancement can be identified on the MIP images by a large difference in signal intensity between organ tissues (blue arrows) and bone structures (red arrows).

Using the HTML page for CT series having eight or more slices, the anatomy field was manually reviewed to determine neural net output accuracy. In cases where the field did not match the respective thumbnail, it was manually corrected in the labeling/QC file. Similarly, if contrast was not observed in the MIP image, the file was edited accordingly. A lack of contrast in the tissue was identified by a large intensity difference between the organ tissue, which increases in intensity in the presence of contrast, and bone structures, which are consistently high-intensity regardless of the presence of contrast (Figure 5).

On HTML pages for NM series, the anatomy field was manually reviewed to determine whether it was correctly classified as AP or PA. The orientation field was reviewed to determine whether the series was correctly classified as an original image or screen capture.

Tagging images for labeling

The goal of the system is to label the optimal image series for reading at each time point. After the labeling/QC file was manually reviewed and edited, automated software updated the label column, which flagged the selected optimal series. The software labeled a single set of required anatomy (Chest–Abdomen–Pelvis for CT, AP and PA for NM) per time point; for these series the label column was set to “Yes”. For CT series, the software first looks to label series having a score of 5, then 4, and then 3, if a higher score is not available. Within these scores, for CT series, the system looked to label full CAP series; if none is available, it separately labeled series that form full CAP coverage when combined. For NM series, the software labeled series having a score of 5, then 3 (if 5 was not available). Within these scores, the script looked to label one AP series and one PA series.

Visit label reconciliation and labeling

Visit labels are uniform time-point identifiers unique to each study (e.g., Screening, Treatment Cycle 4, etc.). Using automated software, visit labels were matched with study dates in the labeling/QC file. Visit numbers (e.g., 1, 2, 3) were then matched with the visit labels in this file. Automated software was used to label all tagged series within our database with their appropriate anatomy, visit number, and visit label. Labeling was accomplished by matching the series unique identifier (series UID) of a tagged series in the .csv file with the series UID of that scan in the database.

Table 3: Image quality criteria reflected in the image QC report and their effect on the overall IQS.

QC criteria	Effect on IQS
Images contain PHI at receipt	IQS decreased to 4
The scanner used at follow-up is different than the scanner used at screening	IQS decreased to 3
A required anatomical scan is missing	IQS decreased to 1
Anatomical coverage is incomplete	IQS decreased to 1
Slice thickness is less than 1 mm	IQS decreased to 2
Slice thickness is greater than 5 mm	IQS decreased to 3
Slice thickness is greater than 10 mm	IQS decreased to 1
CT images are non-contiguous	IQS decreased to 1
CT images are non-contrast	IQS decreased to 3
NM images are screen capture	IQS decreased to 3

Image QC report generation

Using the labeled data in the database, a quality control (QC) report was automatically created for each time point. This QC report is included in the study records, ensures that the relevant sequences were acquired with the correct parameters and appropriate quality, and provides rapid feedback to the imaging site. QC parameters were automatically obtained from the labeled data, including subject ID, study date, and scanner information, along with modality-specific parameters such as slice thickness, contiguity, and contrast/non-contrast for CT series. For NM series, rows, columns, and original/screen capture information were included. An overall image quality score (IQS) was then calculated for each time point and provided to the site as feedback. This IQS had a maximum value of 5 and could be decreased for any of the reasons shown in Table 3. An IQS of 5 was considered excellent quality data; an IQS of 3-4 was considered analyzable data; an IQS of 1-2 was considered partially analyzable data; and an IQS of 0 was considered non-usable data.

RESULTS

A total of 900 patients (9,006 total CT and NM time points, 204,492 total series) have been imported and analyzed to date using the high-throughput pipeline. HTML pages and log files were reviewed in batches of approximately 150 subjects each. Of the total series, 27,630 (13.5%) were CT and 170,860 (83.6%) were NM. The large number of NM series was due to splitting of unusable 3D NM data into an individual series per DICOM image during import. The flowchart below (Figure 6) shows how the series were separated into usable versus not relevant/usable, then into labeled (for reading) versus not labeled. It shows there were many more image series per time point than needed to be labeled, and therefore the system had to sift through many series to identify the subset that should actually be labeled and read.

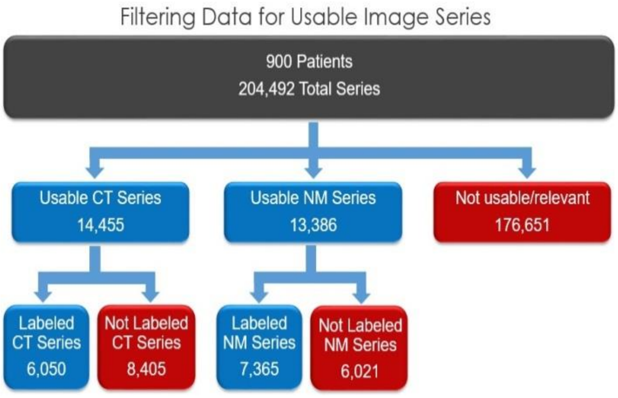


Figure 6: A flowchart describing the process of filtering a large bolus of image series into usable CT and NM series, and then into the optimal series for labeling.

During human review of automatic anatomic classification using the HTML pages, 4.3% of all usable series (5.5% of usable axial CT series and 3.0% of usable NM series) had incorrectly classified anatomy and required correction. Therefore, the neural network was 95.7% accurate in classifying the imaged anatomy. Out of all CT series, 1,126 (4.1%) were non-contrast and therefore required that field to be manually updated. QC reports were successfully created for all 9,006 imported time points.

The majority of incorrectly classified CT series were cases where a partially visible anatomy type was included in the anatomy label, for example when a chest image having partial abdomen coverage was labeled as Chest–Abdomen. Some were also incorrectly classified when image artifacts from surgical hardware or abnormal image intensity ranges were present. The majority of incorrectly classified NM series (AP versus PA, and vice versa) were due to incorrect information in the DICOM header data; for example, a PA scan containing Anterior within its DICOM tag. We observed some series where Rows = 1024 and Columns = 1024 but the series was an original whole body scan (rather than a screen capture) and was adjusted during manual review. However, since the majority of 1024×1024 NM series were screen captures, our default screen capture classification for them was acceptable.

Over a span of four months, all 204,492 image series were checked for PHI, then imported, characterized, labeled, and checked via QC report by a team of two people. The data is currently undergoing quantitative image analysis via lesion marking, to be followed by statistical analysis to achieve the study endpoints.

DISCUSSION

Implementing this high-throughput review process greatly decreased the average number of work-hours spent per time point. As each time point typically requires approximate 30 minutes to manually take through the de-identification, importation, review, and labeling process, processing the 9,006 time points in our dataset would require 4,503 person-hours, equivalent to over two years of work. Using the pipeline described herein, one person, working eight hours a day, was able to process a batch of approximately 1,500 time points (~150 patients) every 1.5 weeks, equivalent to 2.4 minutes per time point. Therefore, we achieved over a ten-fold reduction in the work-hours required to process this dataset.

Previous studies have attempted to describe and improve data intake analysis workflow for clinical trial data in various settings. Meinecke et al focused on the importance of conducting pragmatic trials, where maintaining effectiveness under suboptimal conditions often encountered in clinical practice is necessary to conduct a useful study.[17] Dunn et al have described a custom-built web application for organizing existing and future clinical data rather than the *ad hoc* data capture systems used at many medical centers.[18] Omollo et al have described a workflow for efficient clinical data management using only open-source software for application in low-resource regions.[19] To our knowledge, this paper is the first to describe a high-throughput pipeline for rapid clinical image intake, labeling, and QC using machine learning and automation techniques.

This pipeline can be applied to many types of imaging studies, whenever large datasets make manual review cumbersome and overly time consuming. Many hospitals and institutions use large picture archiving and communication system (PACS) technology to store significant amounts of data that can be queried on individual computers.[20] However, this data is often not searchable beyond typical identifiers, such as patient information and scan date. This classification method could potentially be applied to search PACS images and obtain the highest-quality or most relevant image data needed for a particular research study, rather than the typical method of manually selecting useful image series from queryable data. In turn, this may let PACS systems be used as data input for studies exploring big data and data mining methods that continue to become more prevalent in the realm of clinical oncology.[21]

A non-zero number of series are incorrectly classified by our high-throughput pipeline, approximately 4.3% of overall usable series. It may be possible to further reduce the number of mislabeled CT series with additional neural net training cases, but a small number of false outputs are likely to persist given the heterogeneous nature of image data. Similarly, for NM series, it will be difficult to correctly determine anatomy when its DICOM header data is incorrect, as that is the simplest and often more accurate way to obtain the anatomy. But as demonstrated in this study, the relatively high accuracy of these classification methods combined with using anatomic HTML pages to identify any false classifications remains far less work- and time-intensive than manual labeling methods.

Future improvements to this method may help improve speed and accuracy. One possible avenue for advancement would be implementation of a neural net for classifying contrast and non-contrast CT series, which would remove the need to correct the non-contrast series during manual review. Similarly, by identifying abnormal non-anatomical text features within images, it may be possible to use machine learning techniques to flag series containing PHI, greatly reducing the time required for its removal. It may also be possible to improve the CT anatomy classification criteria shown in Table 1; although our empirically determined criteria worked well to obtain a low error rate, more optimized criteria may be attainable.

In summary, a high-throughput pipeline was applied to greatly decrease the processing time of a large dataset of CT and NM images by accurately classifying image data, performing QC checks, selecting the best series, and tagging images for analysis. This method can be modified and applied to other imaging modalities as needed.

ACKNOWLEDGEMENTS

The authors would like to thank Chris Morton for his assistance in editing this manuscript.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Krishnankutty B, Bellary S, Kumar NBR, Moodahadu LS. Data management in clinical research: An overview. Indian J Pharmacol. 2012;44(2):168-72.
2. Braun R. Systems analysis of high-throughput data. Adv Exp Med Biol. 2014;844:153-87.
3. Yan SF, King FJ, He Y, Caldwell JS, Zhou Y. Learning from the data: Mining of large high-throughput screening databases. J Chem Inf Model. 2006;46(6):2381-95.
4. Sulakhe D, Balasubramanian S, Xie B, et al. High-throughput translational medicine: Challenges and solutions. Adv Exp Med Biol. 2014;799:39-67.
5. Veltri P. Management and analysis of biological and clinical data: How computer science may support biomedical and clinical research. In: Physics Procedia. 2015;62:29-35.
6. Kennan MA, Markauskaite L. Research Data Management Practices: A Snapshot in Time. Int J Digit Curation. 2015;10(2).
7. Cumbaa C, Jurisica I. Automatic classification and pattern discovery in high-throughput protein crystallization trials. J Struct Funct Genomics. 2005;6(2-3):195-202.
8. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228-47.
9. Scher HI, Halabi S, Tannock I, Morris M, Sternberg CN, Carducci MA, et al. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: Recommendations of the Prostate Cancer Clinical Trials Working Group. J Clin Oncol. 2008;26(7):1148-59.
10. Roth HR, Lee CT, Shin H-C, et al. Anatomy-specific classification of medical images using deep convolutional nets. Biomed Imaging (ISBI), 2015 IEEE 12th Int Symp. 2015: 101-104.
11. Ranzato M, Huang FJ, Boureau YL, LeCun Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: Proceedings of the IEEE Computer Society

Conference on Computer Vision and Pattern Recognition; 2007.

12. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res. 2014;15:1929-58.
13. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics. 2017;18.
14. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. Proc 27th Int Conf Mach Learn. 2010;(3):807-14.
15. Kingma DP, Ba JL. Adam: a Method for Stochastic Optimization. Int Conf Learn Represent. 2015. 2015: 1-15.
16. Prokop M, Shin HO, Schanz a, Schaefer-Prokop CM. Use of maximum intensity projections in CT angiography: a basic review. Radiographics. 1997;17(2):433-51.
17. Meinecke AK, Welsing P, Kafatos G. Data collection in pragmatic trials. J Clin Epidemiol. 2017.
18. Dunn WD, Cobb J, Levey AI, Gutman DA. REDLetr: Workflow and tools to support the migration of legacy clinical data capture systems to REDCap. Int J Med Inform. 2016;93:103-10.
19. Omollo R, Ochieng M, Mutinda B, Omollo T, Owiti R, Okeya S, et al. Innovative Approaches to Clinical Data Management in Resource Limited Settings Using Open-Source Technologies. PLoS Negl Trop Dis. 2014;8(9).
20. Mansoori B, Erhard KK, Sunshine JL. Picture Archiving and Communication System (PACS) Implementation, Integration & Benefits in an Integrated Health System. Acad Radiol. 2012;19(2):229-35.
21. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8-17.

Cite this article as: Harris RJ, Teng P, Nagarajan M, Shrestha L, Lu X, Ramakrishna B, et al. High-throughput image labeling and quality control for clinical trials using machine learning. Int J Clin Trials 2018;5(4):161-9.