The Trials and Tribulations of Assembling Large Datasets for Machine Learning Applications

Kirti Magudia, MD, PhD, Brigham & Women's Hospital/Harvard Medical School; Chris P. Bridge, DPhil; Mark Walters; Adam McCarthy, MSc; Mark Michalski, MD; Katherine P. Andriole, PhD, FSIIM; Michael H. Rosenthal, MD, PhD

Background/Problem Being Solved

With vast interest in machine learning applications, more investigators are proposing to assemble large datasets for training, validation and/or analysis using machine learning models for large-scale research studies. Performance of models generally improves with more data so maximal dataset size is desired. Retrieving exams from radiology systems initially appears to be a simple step in a machine learning project. However, many roadblocks may present themselves, leading to significant time delays in gathering the number of exams desired for a given project. We outline our experience in pulling 22,852 CT abdomen/pelvis exams from 2 major academic hospitals with the aim of alerting other investigators to measures that may be taken to save valuable project time.

Interventions

Our aim was to pull all adult outpatient CT abdomen/pelvis exams performed in the Partners Hospital system in 2012. Patients were identified though the Partners Healthcare Research Patient Data Registry. The data provided by this registry included all radiology exams for outpatients that had at least one CT abdomen/pelvis exam in 2012 (1.7 million exams). This data was then limited to all CTs; to patients imaged in the year 2012; to test descriptions of "Abd"; group of exam not "chest," "hdnk" (head and neck), "unclassified," "resp" (respiratory), "lextr" (lower extremity), or "cspin" (cervical spine); to type of patient not "inpatient"; and age between 18 and 99; resulting in 33,182 exams for 23,186 unique patients. We selected the earliest exam in 2012 for each of the included patients to limit our dataset to a single exam per patient.

The roadblocks we faced were in 4 major categories: cohort creation & processing, retrieving DICOM exam files from PACS, data storage and non-recoverable failures. 44% of exams required reformatting of exam accession (ACC) and medical record number (MRN) due to inconsistencies in the formats of these numbers across changes in electronic medical record (EMR) and radiology information systems (RIS). 4% of cases did not have images linked to the expected ACC due inconsistent policies over time; these retrievals were corrected by inclusion of additional exam ACC candidates. Limited archive-side data access and storage mechanisms slowed exam retrieval initially, but new high-throughput systems were successfully implemented. Lastly, there were a total of 51 exams (0.2%) with nonrecoverable failures that were excluded from further analysis. Ultimately, the vast majority of the encountered errors related to inconsistent data policies related to transitions in the underlying EMR and RIS over time (Table 1).

	Problem	Solution	Result
Cohort creation & processing	Mislabeled exams *(i.e interventional and musculoskeletal labeled as abdominal CTs)*	Removed test descriptions of "ablation, fna, biopsy, drainage, guidance, drain, drg, bx, interventional, interv, perc, bone"	22,903 exams remaining
	Inconsistent formatting of medical record numbers (MRNs) and accessions (ACCs)	For one hospital, MRNs were padded with leading zeros to 8 digits. All ACCs before a change in EMR had a leading "A" removed.	MRN and ACCs for 10,089 exams were reformatted
	Some ACCs generated solely for billing with no linked images *(i.e. CT abdomen/pelvis with separate ACCs for abdomen, pelvis and contrast)*	Copies of the underlying databases of both hospital clinical PACS systems were queried with MRN and date to identify all candidate ACCs • CTs with >20 images • Search for exam with body part of "abdomen, GI, GU or body" • If none, search for exams with body part of "pelvis" and then "chest" • If none, expand date range to +/- 4 days	838 exams had different ACC chosen than what was provided from the research database Note: wrong ACC still pulled for 24 exams, which were manually corrected
	Inconsistent linkage of images to ACCs *(i.e. images could be either with the abdomen or pelvis ACCs)*		
Exam retrieval	Slow pull method for one hospital consisted of a Web API pull method that preceded a vendor neutral archive	New method established where radiology IT pushes exams to a DICOM 4CHEE instance (an open source DICOM image management system), which is then transferred to our storage.	Original time estimate to retrieve exams of >1 year. With new method, exams retrieved in 2 weeks
	Push rate from DCM4CHEE instance exceeding write rate to storage, causing crashes	Slowed down push rate and added memory to the server running the DCM4CHEE instance to buffer images as they came in before they were written to storage	No further system crashes during exam retrieval
Data storage	Data storage requirement exceeded available storage in a multiuser system	Transitioned project files to new storage device	Overall delay of 3 weeks
Non-recoverable failures	MRN/ACC mismatches	Exclude from further analysis	17 exams
	Topogram only exam		7 exams
	Missing exam		8 exams
	Corrupted CT data		3 exams
	Non-patient test exam		1 exam
	DICOM encoding errors		15 exams
Final number of exams			**22,852 exams**

Table 1: Problems encountered during cohort creation with corresponding solutions and results.

Outcome

Successful retrieval of 22,852 valid CT abdomen/pelvis exams for analysis by a machine learning algorithm. The overall time required to achieve this outcome was 3 months and at minimum 300 man-hours of time between the primary investigator (a radiologist), a data scientist, and a software engineer. Significantly less time should be required to assemble a similar dataset in the future.

Conclusion
Our experience in retrieving 22,852 valid CT abdomen/pelvis exams identified four major categories of challenges when retrieving large datasets: cohort creation & processing, retrieving DICOM exam files from PACS, data storage and non-recoverable failures. As troubleshooting these issues took three months of project time, we share our experience so that other investigators can anticipate and plan for these challenges.

Statement of Impact
This work describes barriers and solutions for the assembly of large radiology data sets for machine learning applications.

Keywords
cohort creation, data retrieval, data labels, machine learning, PACS