Building a Multi-Center Data Repository of Extensively Labeled Radiology Images of COVID-19 Patients

Gaurav Anand, MD, Yale New Haven Hospital

Introduction

The COVID-19 pandemic continues to cause devastation across the globe, with more than 106 million cases reported worldwide as of February 8, 2021. Although vaccination rates are increasing, the threat of the virus and its variants continues, thus necessitating methods to accurately detect the disease to allow for appropriate treatment.

Hypothesis

We postulate that we can create a repository of labeled radiology images that will facilitate cross-institutional sharing of data and development of AI algorithms to assist in detecting COVID-19. Importantly, the method of labeling these images would be in a manner that would not disrupt workflow.

Methods

A novel method of data collection and storage was used to build this repository (Figure 1). Batched clinical data exported from the EMR were de-identified and imported to the PACS research server. Within the PACS, a COVID-19 icon launched a pre-populated webform containing relevant clinical data and a FHIR questionnaire to tag data associated with the anonymized ID from the PACS research server.

Our system was created to allow PACS to recognize the anonymized ID using a non-reversible hash-based forward mapping system, allowing for the webform to be displayed within the PACS and completed during image annotation. DICOM images labeled with patient demographics, pre-existing conditions, laboratory values, treatment history, and other data were uploaded to the PACS research server using de-identified IDs. The finalized dataset was uploaded to a cloud network made accessible only to collaborating researchers. Future cooperation between PACS and EMR developers may allow for FHIR queries to enable real-time extraction of data from the EMR.

Results

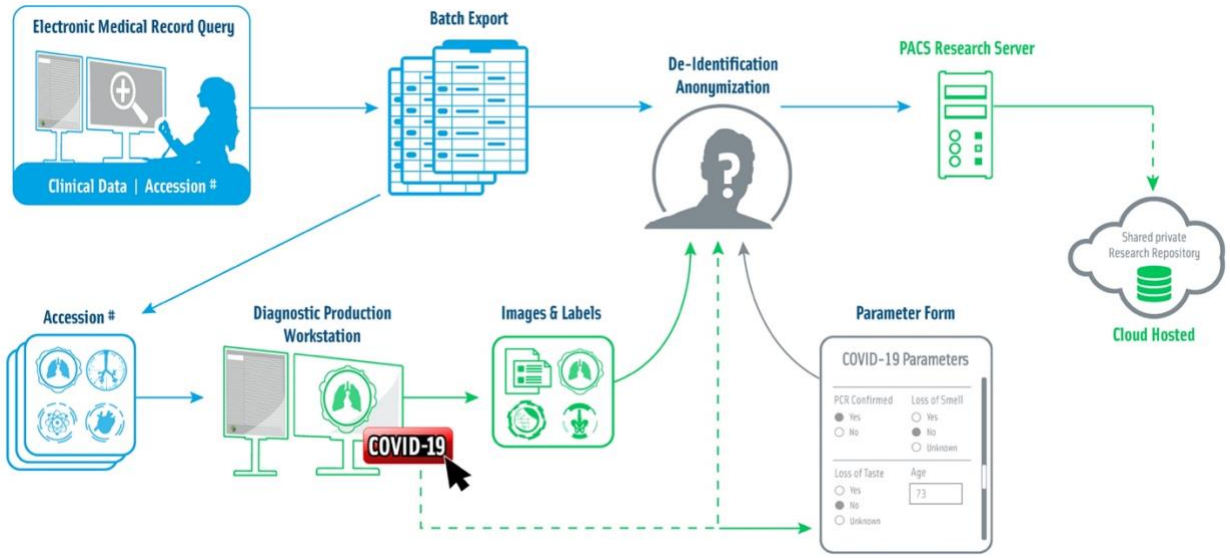
The described method of data collection and storage was implemented at our institution. Taking data from a three-month period at the start of the pandemic, a total of 33946 exams (including

radiographs, computed tomography studies, and ultrasound exams) from 7393 patients were transferred to the PACS research server. These data have proven useful to several research projects ranging from labeling lung imaging to quantifying neurologic complications. Although the focus of this database is currently on COVID-19, this repository will serve as the foundation upon which additional disease states can be addressed.

Conclusion

A shared PACS allows for the creation of a multi-institutional repository of annotated radiology images by sharing processes for de-identification, labeling, storage, and remote access of data. This repository will facilitate cross-institutional research pursuits and the development of AI algorithms that may aid in the fight against COVID-19.

Figure(s)



Keywords

Applications; Artificial Intelligence; Clinical Workflow & Productivity; Emerging Technologies; Imaging Research; Storage