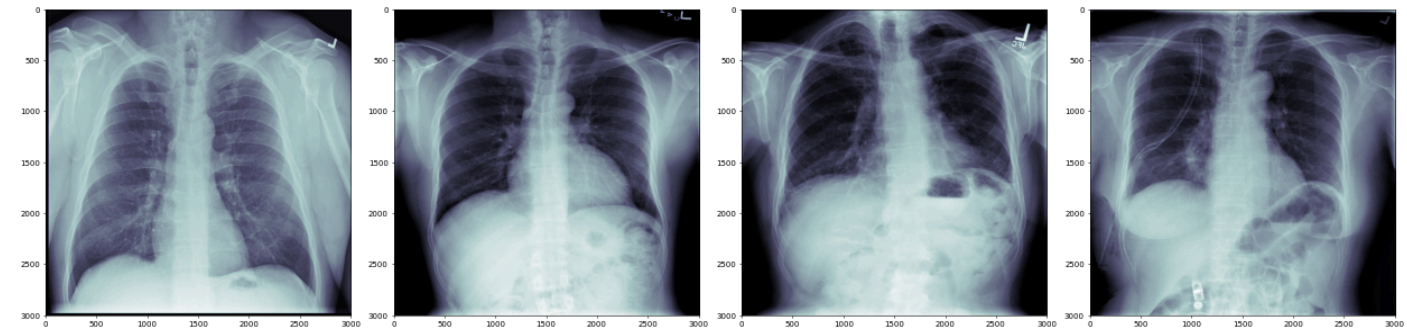
Phase 1: Data Collection

You are a bioinformatics professor working with the Stanford AIMI Center, in charge of a large research lab composed of both bioinformatics and machine learning PhD candidates. Last week, a high-impact journal published a paper introducing a large dataset of COVID-19 data to the public. The data was collected from three clinics (Clinic A, Clinic B, and Clinic C) over the course of the last six months and consists of both chest x-ray (CXR) images and electronic health records (EHR). Your lab decides to embark on two separate machine learning projects leveraging this dataset:

- **Project 1**: a deep learning-based model that will attempt to discern whether or not a given patient has COVID-19 given their chest x-rays, and
- **Project 2**: a logistic regression-based model that will attempt to identify whether or not a patient will need intubation.

The data consists of 30,000 COVID exams across 21,000 unique patients (some patients may be associated with multiple exams). There are 27,000 COVID-negative exams and 3,000 COVID-positive exams, a breakdown of 90% negative cases and 10% positive cases.

Each COVID exam is associated with both a CXR image and EHR data. The CXR image is a 3000 by 3000 pixel image in uncompressed DICOM format. Below is a sample of 4 images:



The EHR data is a large CSV file, where the rows are individual exams and the columns are a mix of exam metadata, observed symptoms, and lab values. The observed symptoms are binary values– for a given exam and a given symptom, the column describing the symptom is 1 if the symptom was observed during the exam and 0 otherwise. Below is a sample of 5 rows:

In [1]:
```
pd.read_csv('COVID_19_sample_data.csv').iloc[5:10]
```

Out[1]:

	pat_deid	intubation_date	IP_admission_date	IP_discharge_date	clinic	birth_date	death_
5	8f9539f2-e6ad-4e00-ad45-4abc2bff2214	NaN	2020-03-04	2020-03-14	Clinic B	1965-11-11	
6	d5dd13c4-c31e-419c-8c02-47e4ca1ac5e2	NaN	2020-03-02	2020-03-20	Clinic B	2018-08-16	
7	91369e11-b944-4132-be0f-af46e880936b	NaN	2020-03-02	2020-03-21	Clinic C	1972-09-22	
8	c70992c9-ff13-467b-9032-1901506edeef	NaN	2020-02-29	2020-03-05	Clinic C	1959-06-17	2020-(
9	c70992c9-ff13-467b-9032-1901506edeef	2020-03-05	2020-03-05	2020-03-12	Clinic B	1959-06-17	2020-(

NOTE: You are also aware of another EHR dataset composed of 40,000 exams from 28,000 patients with varying respiratory illnesses, such as influenza, pneumonia, and viral pneumonia. It has the same set of features as this incoming COVID dataset (list of exam metadata, observable symptoms, and lab values). Keep this in mind– it may come in handy in the future.

In the following quiz, you will answer questions regarding the use of the new publicly available COVID-19 data for the purposes of machine learning, as well as answer questions regarding the approach of the teams working on each of the projects.

In [ ]:
```
```