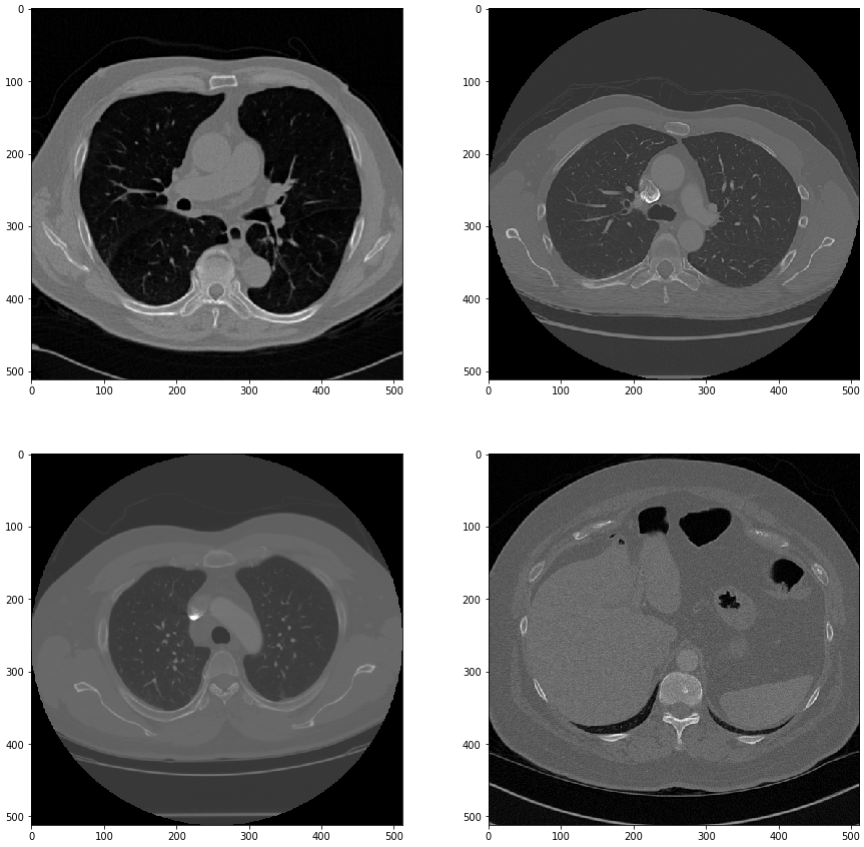
# Group 12 - Cancer Diagnosis in Medical Imaging

## A. EDA

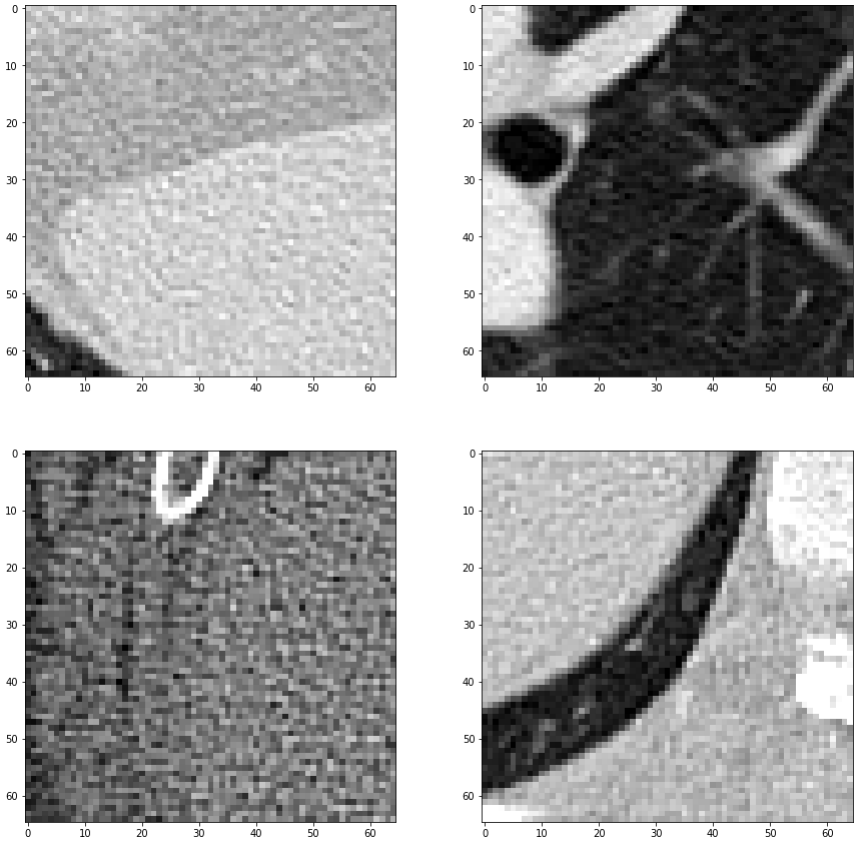
### Sample Images



***Above we see a sample of the images in the dataset. Each sample in our dataset a 3d scan which consists of several images, each representing a slice in the Z direction. The 3d nature of this data is an essential aspect used by human experts, and we expect that it is equally essential to any object detection algorithm.***

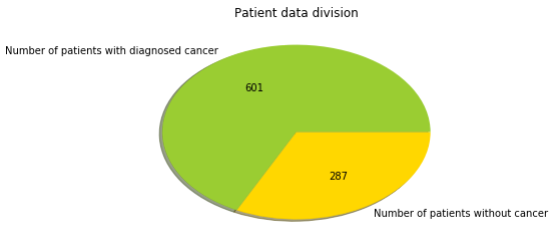
```
IOPub data rate exceeded.
The notebook server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--NotebookApp.iopub_data_rate_limit`.

Current values:
NotebookApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
NotebookApp.rate_limit_window=3.0 (secs)
```

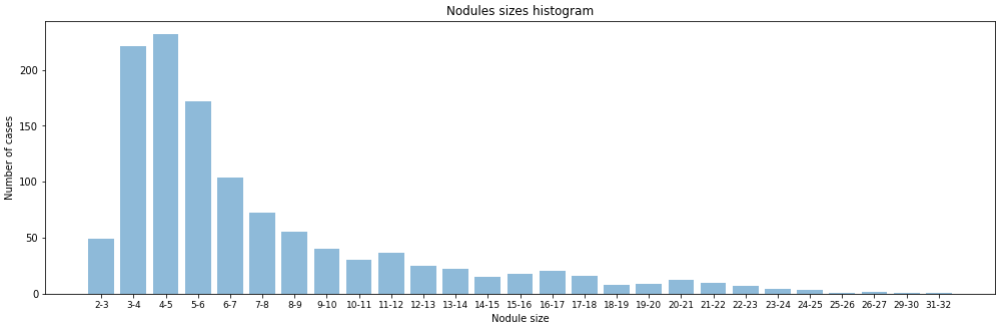


***Above are several sub-sections from slices within our dataset at approximately the scale on which our target features (cancerous nodules) exist. For training 3d convolutional filters to detect such features, our network will recieve 3d cubes of data at this scale.***

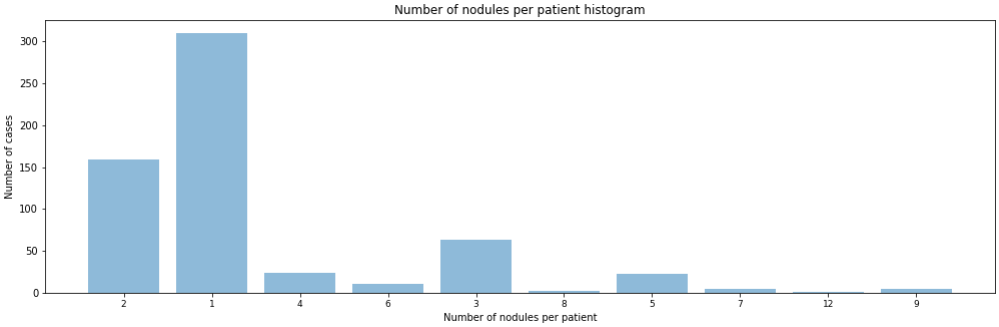
### Metadata about our sample




While many medical imaging datasets contain only samples of patients diagnosed with the affliction under study, our dataset does contain a significant portion of healthy patients which should expose the model to, and allow us to test on, data that better reflects what a model would see if used in actual patient care.




Most nodules are 3-6mm, but in order to capture the largest we will need to be sure that our 3d chunks capture those up to 30mm.




### B. Revised Project Statement / Goals

After reviewing the 2017 work of Baker, Kilpatrick, and Chaudhry (http://cs231n.stanford.edu/reports/2017/pdfs/515.pdf (http://cs231n.stanford.edu/reports/2017/pdfs/515.pdf)), we believe that the most significant contribution may be made by improving on the use of 3d conv nets to identify cancerous nodules. While Baker, et al. were unable to achieve compelling results with 3d conv nets, we expect that more modern techniques may yield significantly better accuracy.

Noting that Baker, et al. achieved their strongest results by adapting 2d network pre-trained in ImageNet, we intend to pre-train our 3d convolutional filters within an autoencoder. While our classification task will be limited to nodule candidate areas, the pre-training will be conducted over the entire dataset in order to encourage the filters to learn diverse features. This will also provide us with a sense of the representational capacity needed for this 3d data set, an important task itself as most computer vision is presently done in two dimensions and our intuition likely needs significant recallibration for three. We also plan to experiment with GANs to see if they might also contribute to better accuracy in three dimensions.

The classification task will be performed on cubic chunks of the source images. Chunks are taken of the regions surrounding 'candidates' identified in the dataset. Each candidate is a region that may, but does not neccesarily, contain a cancerous nodule. The classification is binary, but the classes are heavily imbbalanced to favor non-cancer. We will use data augmentation methods, ie rotating / flipping / reversing our cubic chunks, to produce more varied samples.

For evaluating accuracy, we are interested first in the classification ability of the model on 'candidate chunks', that is the areas which have been identified in the dataset as needing classification. We will examine the ROC curve for our model in order to assess this. Afterward, we are interested in the models ability to locate cancerous areas when presented with an entire scan. While numerous object detection and localization algorithms exist, a simple technique such as sliding-windows is sufficient for evaluating the ability of our model to independently detect cancerous nodules.

Out[2]: Hide code.