Background

The diagnosis and treatment of lung cancer has been drastically improved by newer imaging methods like PET-CT which generate large number of images where single spots can drastically influence the diagnosis and treatment. For physicians, this means a long time must be spent carefully reading images and looking at dozens of different regions carefully to check for the possibility of aggressive disease. CADx, machine learning and workflow automation have demonstrated the potential to radically reduce the reading time without sacrificing quality in a number of different medical areas ranging from pathology for radiology. These cases without exception have had large numbers of patients (>1000), while lung cancer studies with PET-CT typically have fewer than 50-60 patients [1]. We undertake a challenge to perform such a large scale study in Nuclear Medicine in order to build the basis for developing better, smarter automated approaches. We present the work we have done on 300 (of the total 2000 patients) towards a computer aided staging of Non-Small Cell Lung Cancer (NSCLC) with particular emphasis on gathering large studies, consistently preprocessing PET-CT image data, and quickly annotating large patient groups.

Evaluation

The study consists of 300 patients from several different hospitals split evenly between those with confirmed NSCLC and those without (as a control group).

To collect the data, we had to build system to bring together of disparate databases: PACS, RIS, and spreadsheets from Oncology. The project, called PACSCrawler, we have made open-source [2] and enables scalable collection and searching of large amounts of meta- and text data. Once running, this enabled us to find a suitable cohort of patients for such a study with ease.

Once the image data is collected, the data must be standardized for further steps. A preprocessing pipeline has been built to ensure all the images have standardized resolution, scaling, reconstruction and normalization. Finally, once the data have been standardized, they anonymized and packaged as 3D arrays in HDF5 to allow for standard machine learning tools like Python, R, Matlab to easily access and further process the data. Using SimpleITK and Python, we have built up a pipeline and have demonstrated the results of the preprocessing are shown on Cancer Imaging Archive Soft-tissue Sarcoma [3] data in the form of an Open Data challenge on Kaggle [4].

After preprocessing is completed, the expert knowledge of the radiologist and nuclear medicine physician need to be extracted. We do this in two parts. The first is through careful annotation of the individual images. As none of the tools had the flexibility or features needed out of the box, we extended the 3D Slicer [5] project to handle the needs of PET-CT data and NSCLC. The
annotations take approximately 20-30 minutes per patient. The second part of this structuring of knowledge is by recording a classification from the physician. As the most relevant stating for the radiologist and the tumor board is the TNM-staging, 275 different TNM stagings (categories) ranging from no signs of cancer: T0 N0 N0 to the extreme case: T4 N3 M1b are recorded (one for each patient). The result is a structured library of patients, identified lesions, and official stagings. These can then be used to train machine learning approaches for finding and classifying lesions, as well eventually for end-to-end staging of patients.

**Figure 1**

The figure shows the output of the preprocessing pipeline (standardized CT and PET images in the first two columns) and the result of 30 minutes of annotation in the third column.

**Figure 2**

The web-interface for text-based searches (left panel), view radiology reports (bottom right), and see summary overviews of different factors for a study group: distributions of staging, age, or classification code (upper right).
The annotation tool with fusion PETCT images is shown above. The 41 different labels which can be chosen are shown in the first two rows of the toolbar (single click to mark a new region), the other settings for quickly adjusting window and viewing settings are all available on the same screen.

The figure shows the annotated fusion slices produced by the radiologist on the left and the 3D renderings of the lesions overlaid on the skeletal structure on the right.
Discussion

The realization of such a large-scale study presented many challenges in the infrastructure and organization aspects. In particular, the systems in place in a clinical environment are poorly suited for ‘Big Data’ and radiomics style approaches and require significant modifications. Furthermore, the diversity of data present when collaborating between not only different hospitals but different divisions within a hospital (Oncology, Radiology, Nuclear Medicine).

The preprocessing pipeline was a challenge since the CT and PET images came from several different machines from different manufacturers. This lead to a very large heterogeneity of the input data where even naming conventions were rarely upheld. Particularly with PET images, we found it is critical that the factors such as dose, patient weight, and attenuation correction are properly, consistently handled.

Conclusion

While the challenges in Nuclear Medicine are not trivial to resolve, the potential for large scale analysis is present. In particular, the ability to cleverly combine multiple modalities of information and extract information from cross-modal features and textures present exciting opportunities for further algorithmic exploration and development.

References

2. PACSCrawler: https://github.com/joshy/meta
5. https://www.youtube.com/watch?v=4KUDniepg74

Keywords

NSCLC, PET-CT, Multicenter Studies, Machine Learning, Deep Learning, Image Analysis, Texture Analysis, Feature Extraction