**Background: Briefly describe previous findings or observations that provide the background leading to this proposal.**

The proposal is inspired by the work of Shin et al [1], on representing semantic interactions between radiology images and clinical reports obtained from a large-scale hospital Picture Archiving and Communication Systems (PACS) database. They used Latent Dirichlet Allocation (LDA), topic modeling method, to predict the semantic topics and keywords for radiology image slices and performed transfer learning on pre-trained Convolution Neural Network (CNN) on natural images to build an image to topic mapping. Further work done by same group [2] used radiology dataset of chest x-rays and their clinical reports to train a CNN to efficiently detect a disease from an image and use the deep features to further train a RNN to describe the context of a detected diseases. Further in [3], they present a Looped Deep Pseudo-task Optimization (LDPO) approach to automatically discover visually coherent and clinically semantic clusters. They learn and finetune pre-trained CNN model using pseudo-tasks labels and iterate this process until convergence to latent true image categories obtain by performing LDA on clinical reports. Another work [4], proposed a feature learning approach that separately clusters 3D lung CT data based on visual features and terms in radiology reports based on LDA and then establish links between the two types of clusters. [5] presents a multi-task CNN approach to predicted categorical Breast Imaging Reporting and Data System (BI-RADS) descriptors for breast lesions.

[1] Shin, H.-C., Lu, L., Kim, L., Seff, A., Yao, J., Summers, R. M., 2015. Interleaved text/image deep mining on a very large-scale radiology database. In: Computer Vision and Pattern Recognition. pp. 1090–1099

[2] Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R. M. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. arXiv:1603.08486

[3] Wang, X., Lu, L., Shin, H.-c., Kim, L., Nogues, I., Yao, J., Summers, R. Unsupervised category discovery via looped deep pseudo-task optimization using a large-scale radiology image database. arXiv:1603.07965.

[4] J. Hofmanninger, M. Krenn, M. Holzer, T. Schlegl, H. Prosch, G. Langs. 2016 Unsupervised Identification of Clinically Relevant Clusters in Routine Imaging Data, In: MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I. Springer, pp. 192--200

[5] Kisilev, P., Sason, E., Barkan, E., Hashoul, S., 2016. Medical image description using multi-task-loss CNN. In: International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. Springer, pp. 121–129

**Describe the statistical approaches that will be used to analyze the study data.**

We propose a non-parametric approach for characterizing heterogeneous diseases and generating a textual summary of the observations. Following the approach described in [1], each subject is represented by a collection of image descriptors extracted from different regions. We proposed to learn these image descriptors using a deep CNN based on the method presented in [2]. The image descriptors are then modeled to non-parametrically estimate divergences between the densities (which correspond to individual patients) from image data instead of directly parametrizing the probability densities. The proposed estimator is based on a nearest neighbor graph. The graph enables us to map the predictions of the clinical measurements back to the anatomical domain. This information is used along with clinical reports to formulate meaningful textual summary of the predictions. The clinical reports are usually short text. The sparsity of content in short text leads to poor performance of conventional topic modelling methods like LDA. We proposed to extend the existing Biterm topic Model (BTM)[3] for efficient topic learning and adapt it for clinical report dataset.

[1] J. Schabdach, W. M. Wells, M. Cho, K.N. Batmanghelich, 2017 A Likelihood-Free Approach for Characterizing Heterogeneous Diseases in Large-Scale Studies

[2] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Paczos, R. Salakhutdinov, A. Smola, 2017. Deep Sets. arXiv:1703.06114

[3] Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12), 2928-2941.