**Optimization of after-care for total hip replacement based on random forest semi-supervised active learning -- proposal**

1. Introduction

Total hip replacement as a highly successful surgery satisficed by patients suffered from hip pain. However, a 20-year follow-up for total hip replacement patients shows that there are 10.7% possibility of revision 20 years after the surgery. [1] As pressures on the NHS grow, it is increasingly difficult to see patients in person either at the hospital or through a home visit. To maximize the usage of limited number of beds in hospital and help doctor to decide which patients are mostly affected and need to be cared priority, this project will use the combination of semi-supervised learning and active learning to distinguish whether, in which ways and when the patients need follow-up.



1. Data



The data in this project is a unique database of records about the hip replacement operations, after-care and recovery of patients over 32 years from Orthopaedical Department at the Royal Devon and Exeter Hospital. The total size of data is which contains rows and columns. Given that it is a large dataset contained 32 years of records and some patients done the total hip replacement since 1990s, we will focus on more recent data for drift and shift of surgical method and material of replaced hip. It is worthily noticed that there is no target label in the dataset, so we need to use active learning which cooperates with doctors to add some label for training and testing.



1. Problems & corresponding methodologies



* 1. Does a patient need a follow-up appointment?

As initial problem, whether a patient need to be followed-up requires building a classifier to justify the status as ‘necessary’ or ‘not necessary’ by binary label. Given that the dataset contains limited labels, a semi-supervised classifier [2] will be used for those unlabeled data for its economical and functional features. Random forest classifiers are desirable here, because of their robustness, stability and ability to handle categorical variables. For those ambiguous classifications, clinicians would check their real categories. Then the checked data will add into the training data. Through this process, the classifier will be retrained, and this loop will continue until satisfactory accuracy score is obtained. This procedure is also be called active learning. Using these labeled data, a second classifier will be trained to distinguish whether a patient need to be followed-up. A patient that had no necessary appointments should be classed as not requiring follow-up.

* 1. Should a follow-up appointment be in-person or via telephone consultation?

After knowing which person required follow-up appointment, another classifier need to be trained based on the labels we generated in the previous part. The semi-learning still will be used, because there are limited labeled data for appointment as well. Random forest and will also be used here (?). The ambiguous classification will be distinguished by clinicians and will be labeled. These data then will add to training data set looping until a good accuracy score is got which means active learning will continuously be used. Using this classifier, patients will be classified whether they need in person appointment or telephone consultation.



* 1. When should a patient attend a follow-up appointment?

In this part, how many days it needs to take after the total hip replacement surgery will be predicted. Regression algorithm will be used here. The specific algorithm will depend on the distribution of dataset or using random forest (?). In this part, overall dataset with follow-up requirement will be used and the regression would predict for both in-person appointment and telephone consultation.



1. Potential risks



Because of the dataset used is generated in the real world, there would be some limitations probably resulting in invalidated result. First and the foremost, the dataset used for this project actually doesn’t have any labels. The semi-supervised learning has probability will not work. Besides, because of the size of dataset is exceptionally large and computational ability for computer used in the project is limited, the project might would facing computational risk. What’s more, the dataset contains the records of patients in last 32 years, but surgical condition or treatment method for surgical aftercare might change from last 32 years. Since it’s historical data, it might be changed in some situation such as copied data to new system in system update. Based on these reasons, we would concentrate on more recent data as much as possible and that would depend on if there are sufficient ones.



[1] L. Neumann, Knude G. F., K. Harry S., Long-term results of Charnley total hip replacement review of 92 patients at 15 to 20 years, March 1994, The Journal of Bone and Joint Surgery, pp. 245-251, avaliable from: https://doi.org/10.1302/0301-620X.76B2.8113285

[2] C. Leistner, A. Saffari, J. Santner and H. Bischof, "Semi-Supervised Random Forests," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 506-513, doi: 10.1109/ICCV.2009.5459198.  
  
[3] S. Wagner, T. Hastie, B. Efron, “Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife,” Journal of Macine Learning Research, 2014, pp. 1625-1651.