論　文　概　要

論 文 題 目 ： Utilizing machine learning for detecting patients with IgA nephropathy from computerized medical bill database

（電子化レセプトデータベースにおける IgA 腎症罹患者特定のための機械学習の活用）

指導教員：

人間総合科学研究科　疾患制御医学専攻

山縣邦弘　教授

所　　属：筑波大学大学院人間総合科学研究科

疾患制御医学専攻

氏　　名：角田　亮也

目　　的:

Chronic kidney disease (CKD) is a medical condition that constitutes a worldwide health burden. More than 490 million people suffer from CKD that leads to dialysis or kidney transplantation or severe cardiovascular events. Immunoglobulin A Nephropathy (IgAN), the most common type of glomerulonephritis that leads patients to CKD, is one of the biggest concerns in the field of nephrology. However, there has been few discoveries of innovative therapies for underlying cause of the disease. One reason is the difficulty in constructing the cohort of IgAN with enough number of patients and enough observation period due to its long-term natural history. Using bills database under the universal healthcare system in Japan has a potential to make a breakthrough by constructing the largest cohort of the world. However, diagnosis codes aimed at billing should not be directly used for clinical research because of the risk of misdiagnosis. To solve this problem, constructing a method to detecting patients with IgAN from bills data is needed. Against this background, the aim of the study is to investigate the efficacy of machine learning technology for predicting true diagnosis of IgAN on bills data.

対象と方法:

This is a descriptive, analytic study that targets 3,743 patients consulting the Department of Nephrology, University of Tsukuba Hospital at least once from January 1, 2013, to July 31, 2019. Medical records and medical bills of the participants are extracted from the data server of the medical ordering system of the Hospital. From all participants, 437 patients were labeled to have been diagnosed IgAN through a review of medical records. Manual analysis and analysis using machine learning were performed. For manual analysis, the sensitivity and the specificity for each diagnostic criterion was described. For machine learning, datasets were processed in three patterns by the compilation of similar medications and diagnosis codes and given to the XGboost program in the way of five-fold cross-validation. We compared manual analysis with the three patterns of machine learning in the performance for the diagnosis of IgAN.

結　　果:

Diagnosis codes of IgAN were only provided for about half of patients. With manual analysis, the three best positive predictive values are shown in the criteria “any N02x or N03x code combined with tonsil treatment,” “any N02x or N03x code combined with both needle biopsy with immunostaining/immunofluorescence and

corticosteroid infusion," and "any N02x or N03x code combined with both two or more measurements of serum IgA and corticosteroid infusion." The positive predictive values for each were 1, 0.7031, and 0.6825, respectively. The maximal sensitivity of 0.6316 was shown in the criterion of "measurement of serum IgA." Both the specificity and the sensitivity of manually created criteria that consists of the recommended examinations and treatments in Japanese guideline for IgAN were lower than 0.8. In contrast, the machine learning process yielded better results in terms of the area under the curve of ROC curve. Especially, mean AUC values over 0.9 were achieved by preprocessing of diagnosis codes and medication codes in a clinical viewpoint.

考　　察:

Although high specificities were shown in some manually constructed criteria, sensitivities were extremely low to be practically used in all criteria. This is the first report of successful challenge to improve the possibility of detecting patients with IgAN from medical bills data only, using machine learning technology in appropriately organized forms. However, problems of external validity due to local factors of the institution and the region remains unclear. The change in codes on the time course also need to be solved to apply this technology for bills data from other institutions.

結　　論:

Compared to manual analysis, applying machine learning technology achieved high performance for detecting patients with IgAN. Preprocessing of dataset in a clinical viewpoint is also important to enhance the efficacy of learning. Improvement and validation using data from another institution is needed.