Review of *Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer*

Grace Yi Chen

The authors built a convolutional neural network (CNN) prediction model for 1-year incidence of nonmelanoma skin cancer (NMSC) based on non-imaging information. The training of the model used 1829 patients first diagnosed with NMSC and 7665 random controls without cancer. The outcome of the model is 1-year incidence of NMSC, and the covariates of the model are time-varying covariates such as 3-year sequential diagnostic and drug prescription information, and patients' baseline variables such as age and sex. In terms of the performance of this model, the AUROC is 0.89 with high sensitivity and specificity. Finally, the model identified chronic comorbidities and medications such as trazodone, acarbose, nonsteroidal anti-inflammatory drugs, and thiazide diuretics to be the most discriminative features in the model.

I think the authors successfully developed a powerful prediction model for NMSC. We know some of the common risk factors for NMSC are UV radiation, ionizing radiation and family history of skin cancer. However, previous studies showed that traditional prediction models using these risk factors as covariates could not achieve a good prediction. More powerful machine learning and deep learning models utilizing electronic medical records are needed. So, the authors used clinical and drug prescription information as a time dependent variables and use CNN to build the prediction model. CNN is usually an efficient tool for image data classification as it considers spatial correlation. The clinical and drug prescription data are time dependent and utilizing the shared weights in CNN makes sense here. In addition, making inference using machine learning variables is hard. The study used stepwise feature selection method to determine the impact of covariates on the AUROC. I think this is a good attempt since the exact coefficients are meaningless, but we could interpret their effect based on AUROC. The study also mentioned some limitations. For example, it did not include some of the known risk factors like exposure to sunlight, family history and genetic parameters. That would be interesting to see whether these factors could improve the CNN model. Also, the authors discussed how the risk probability score thresholds are determined. They did not use an arbitrary threshold 0.5 to classify disease or non-disease. Instead, the threshold was determined based on obtaining the maximal sum of specificity and sensitivity. I am not sure if this is because of the imbalance of sample size in the data and maybe we could investigate more into it.

Questions:

1. The data used in this model includes an imbalance number of cases and controls. I am wondering how two groups are matched as from table 1 there are differences in age and annual drug prescription distributions?
2. Figure 1 shows how time varying covariates are included in the CNN model. I am wondering how time-independent variables are included in the model?