

**Prediction of chronic kidney disease in people with diabetes**

N. Afshar<sup>1</sup>, M. Andorra<sup>2</sup>, S. Chittajallu<sup>1</sup>, T. Huschto<sup>3</sup>, V. Babinsky<sup>4</sup>, H. Mikulski<sup>5</sup>, H. Koenig<sup>3</sup>, C. Ringemann<sup>3</sup>, J. Odegard<sup>6</sup>, I. Singh<sup>7</sup>;

<sup>1</sup>Math Algorithms and Data Analytics, Roche Diabetes Care, Inc., Indianapolis, USA, <sup>2</sup>Math Algorithms and Data Analytics, Roche Diabetes Care, Sant Cugat, Spain, <sup>3</sup>Math Algorithms and Data Analytics, Roche Diabetes Care, Mannheim, Germany, <sup>4</sup>Medical Affairs, Roche Diabetes Care, Vienna, Austria, <sup>5</sup>Medical Affairs, Roche Diabetes Care, Sant Cugat, Spain, <sup>6</sup>Strategy and Customer Solutions, Roche Diabetes Care, Inc., Indianapolis, USA, <sup>7</sup>Professional Solutions Franchise, Roche Diabetes Care, Inc., Indianapolis, USA.

**Background and aims:** People with diabetes in general have a higher risk of developing Chronic Kidney Disease (CKD). The early detection of CKD in people with diabetes helps their health care providers to adjust their therapy to slow down the disease progression to higher stages, prevent complications, and reduce cardiovascular-related conditions. It also takes a great financial burden off the insurance companies in the long run by way of future treatment costs. We have developed a model that uses the historical data from the past 2 years of people with diabetes (PwD) and predicts the chance of them having CKD in the next 3 years.

**Materials and methods:** Our model employed the XGBoost (eXtreme Gradient Boosting) algorithm. It used more than 860,000 PwD data from the IBM EHR (Electronic Health Records) database for training and optimizing the model parameters. It was then independently verified using more than 500,000 PwD data from the CPRD (Clinical Praxis Research Datalink) database from the UK, and close to 140,000 PwD data from the INPC (Indiana Network for Patient Care) database from the US. We considered an incident of CKD based on the appearance of a relevant ICD code. The major benefit of the XGBoost algorithm in the context of risk prediction based on EHR data is that there is no need for imputation of the missing values.

**Results:** We evaluated the performance of the Roche XGBoost model for CKD in terms of the area under the receiver operating characteristic curve (AUC). While we obtain comparable results for the US-based data sets (prediction model:  $AUC_{\text{Explorys}}=0.83$ ,  $AUC_{\text{INPC}}=0.84$ ), we observe a performance decrease for the validation on CPRD data ( $AUC_{\text{CPRD}}=0.74$ ). However, on all data sources our XGBoost model shows a superior performance when compared to appropriate benchmark algorithms from literature.

**Conclusion:** The XGBoost model for CKD disease prediction has a superior performance over benchmark models from literature.

*Disclosure:* N. Afshar: None.