**Objective** To identify existing prediction models for the risk of development of type 2 diabetes and to externally validate them in a large independent cohort.

**Data sources** Practice Fusion is America's fastest growing Electronic Health Record (EHR) community, with more than 170,000 medical professional users treating 34 million patients in all 50 states. Practice Fusion’s EHR-driven research dataset is used to detect disease outbreaks, identify dangerous drug interactions and compare the effectiveness of competing treatments.

**Design**Performance of the models was assessed in terms of discrimination (C statistic) and calibration (calibration plots and Hosmer-Lemeshow test).The validation study was a prospective cohort study, with a case cohort study in a random subcohort.

Predict the probability that each person has a diagnosis of Type 2 Diabetes Mellitus. Predictions are evaluated using the log loss metric.

Log loss is defined as:

log loss=−1N∑i=1Nyilog(yi^)+(1−yi)log(1−yi^),log loss=−1N∑i=1Nyilog⁡(yi^)+(1−yi)log⁡(1−yi^),

where NN is the number of patients, loglog is the natural logarithm, yi^yi^ is the posterior probability that the ithith patient has diabetes, and yiyi is the ground truth (yi=1yi=1 means the patient has diabetes, yi=0yi=0 means that he does not).

https://www.kaggle.com/c/pf2012-diabetes#datasetpreparation