**PROBAST**

Study:

Enhancing self-management in type 1 diabetes with wearablesand deep learning

Step 2: Type of prediction study

**Is the study a diagnostic or a prognostic study?**

Prognostic

**Is the study a development only, development and validation or validation only study?**

**Development only**

**What is the model of interest?**

Recurrent neural network

**What is the outcome of interest?**

Hypoglycemia detection

Step 3: Assess risk of bias

**Domain 1: Participants**

**Describe the sources of data and criteria for participant selection**

Type 1 Diabetes patients. No eligibility criteria

**1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?**

**Y**

**1.2 Were all inclusions and exclusions of participants appropriate?**

Y

**Risk of bias introduced by selection of participants:**

**Low**

**Rationale of bias rating**

No further eligibility criteria specified

**Domain 2: Predictors**

**List and describe predictors included in the final model, e.g. definition and timing of assessment**

Participants were asked to log daily events such as, insulin doses in units, meal macronutrient composition in grams, alcohol intake in units, stress, illness, and exercise in the mySugr smartphone app, which are used to develop the input features of glucose prediction models.

The measured physiological variables applied to theregression analysis include the mean values, standard deviation, range,and maximum and minimum differential values of EDA, IBI, acceleration,and skin temperature signals.

**2.1 Were predictors defined and assessed in a similar way for all participants?**

Y

**2.2 Were predictor assessments made without knowledge of outcome data?**

Y

**2.3 Are all predictors available at the time the model intended to be used?**

Y

**Risk of bias introduced by predictors or their assessment**

Low

**Rationale of bias rating**

Independent and applicable predictors.

**Domain 3: Outcome**

**Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:**

Hypoglycemia as determined by CGM sensor

**3.1 Was the outcome determined appropriately?**

Y

**3.2 Was a pre-specified or standard outcome definition used?**

Y

**3.3 Were predictors excluded from the outcome definition?**

Y

**3.4 Was the outcome defined and determined in a similar way for all participants?**

Y

**3.5 Was the outcome determined without knowledge of predictor information?**

Y

**3.6 Was the time interval between predictor assessment and outcome determination appropriate?**

Y

**Risk of bias introduced by the outcome or its determination**

Low

**Rationale of bias rating**

Outcome is independent and determined with the same CGM device for each patient.

**Domain 4: Analysis**

**Describe number of participants, number of candidate predictors, outcome events and events per candidate predictor**

12 patients. We collected a median (IQR) of 1113.5 (1059.0–1184.0) and 832.5(733.0–953.0) hours of glucose data and sensor wristband data, respectively, and received a total of 5767 daily entries with a median (IQR) of 396 (237–732.3) interactions.

**Describe how the model was developed, predictor selection and risk group definition**

We obtained a total of 20 features from the pre-processed multimodal data (Supplementary Table 3). There are some inevitable errors in the sensor data, e.g., compression artifacts, signal loss, and sensor calibration. To this end, we performed feature selection in the following steps. First, we analyzed the missing fraction of CGM and wristband measurements to identify the quality of features. The median value of the missing percentages of CGM and wristband data are 3.02% and 23.05%, respectively, which are reasonable since the wristband needs to be charged for around 4–5 hours every day. We linearly interpolated the gaps that occurred in the middle of input sequences and extrapolated the gaps at the tail to guarantee that future information is not involved in current predictions. Then, min-max normalization was adopted to scale the selected features to. Finally, we performed collinearity analysis, considering correlated bias is prone to degrade the stability and interpretability of machine learning models62. We noted that features derived from the same measurement exhibited strong a correlation with each other. Hence, each time we retained one feature in IBI or EDA feature group and selected the best combination according to the error scores that summed up RMSE results for the four prediction horizons in model validation.

After pre-processing the features, we developed an attention-based RNN with GRUs for glucose prediction and hypo- and hyperglycemia detection. The multivariate input data for the RNN model were selected according to validation performance, which include CGM, carbohydrate amount, insulin bolus, time index, IBIs, and SCRs.

**Describe whether and how the model was validated, either internally (cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants)**

Considering the personalized models are provided to the T1D subjects at a midterm clinical visit (Fig. 1), we divided the data of each subject into a training set and a testing set that include the first 50% data and the remaining 50% data, respectively. The last 20% data of each training set were used as a hold-out personalized validation set

**Describe the performance measures of the model, e.g. calibration, discrimination, classification, net benefit, and whether they were adjusted for optimism**

ACC, SEN, SPE, PRE, MCC

**Describe any participants who were excluded from the analysis**

None

**Describe missing data on predictors and outcomes as well as methods used for missing data**

We linearly interpolated the gaps that occurred in the middle of input sequences and extrapolated the gaps at the tail to guarantee that future information is not involved in current predictions

**4.1 Were there a reasonable number of participants with the outcome?**

Y

**4.2 Were continuous and categorical predictors handled appropriately?**

Y

**4.3 Were all enrolled participants included in the analysis?**

Y

**4.4 Were participants with missing data handled appropriately?**

Y

**4.5 Was selection of predictors based on univariable analysis avoided?**

Y

**4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls)**

**accounted for appropriately?**

Y

**4.7 Were relevant model performance measures evaluated appropriately?**

Y

**4.8 Were model overfitting and optimism in model performance accounted for?**

Y

**4.9 Do predictors and their assigned weights in the final model correspond to the results**

**from multivariable analysis?**

Y

**Risk of bias introduced by the analysis**

Low

**Rationale of bias rating**

Due to time series there are enough data. Proper handling of missing data and proper validation approach for time series.

**Overall Risk of bias**

Low