**PROBAST**

Study:

Artificial intelligence based real-time prediction of imminent heart failure hospitalisation in patients undergoing non-invasive telemedicine

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?**

Random forest

**What is the outcome of interest?**

Future heart failure hospitalization

Step 3: Assess risk of bias

**Domain 1: Participants**

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

This analysis was conducted on the data of the TIM-HF2trial, a randomised, multi-centre trial assessing the benefits of a structured remote patient management (RPM) programme. In brief, TIM-HF2 was conducted in Germany between 2013 and 2018 and included 1,571 patients with a history of HF, New York Heart Association (NYHA)class II or III and a HF hospitalisation not longer than one year prior to randomisation, regardless of the left ventricular ejection fraction (LVEF).

Patients were randomized to either RPM + usual care or to usual care only, and followed for 12 months. All patients underwent quarterly out-patient visits consisting of medical history, physical examination, collection of blood samples for biomarkers and assessment of concomitant treatments. Patients assigned to RPM were equipped with and trained in the use of a home telemonitoring system, which transmitted body weight, blood pressure, heart rate, ECG recording, peripheral capillary oxygen saturation, and self-rated well-being on a scale from one through five to a telemedical centre (TMC) on a daily basis.

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

PY

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

Y

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

Low

**Rationale of bias rating**

The dataset was made up of two different study arms where patients were treated differently. One group was monitored while the other was not. Groups were distinguished in the analysis

**Domain 2: Predictors**

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

we considered 18 variables (5 binary, 13 numerical) resulting from daily data transmissions, including ECG characteristics, blood pressure, oxygen saturation, weight, and self-rated well-being (Table 2). Additionally, we considered the baseline risk variable, and whether a previous hospitalization due to HF occurred within 30 days prior to data transmission. This resulted in a total of 20 candidate predictors that were considered for the model.

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

U

**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**

Unclear

**Rationale of bias rating**

Due to the two different study arms it is possible that predictors were assessed differently between the patients.

**Domain 3: Outcome**

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

**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?**

PN

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

Y

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

High

**Rationale of bias rating**

It is likely that information of predictors were used to determine whether patients should be hospitalized or not.

**Domain 4: Analysis**

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

The model for the baseline risk variable was validated in the hold-out validation set of 385 patients.

The risk for HF hospitalisation within the 7 days following data transmission was modelled via a Multilayer Perceptron(MLP) (29) using the daily data transmissions of 570 patients from the RPM group over a one-year study period.

During ML modelling on the daily transmitted data, we faced two key challenges: a severe imbalance between patients with and without HF hospitalisations in the next seven days (training prevalence 0.6%), and a lack of heterogeneity in the training set due to the ∼180,000 data transmissions stemming from only 570 individuals.

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

The risk for HF hospitalisation within the 7 days following data transmission was modelled via a Multilayer Perceptron(MLP) (29) using the daily data transmissions of 570 patients from the RPM group over a one-year study period.

**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)**

The main ML model for prediction of unplanned HF hospitalisation within the following seven days was validated in the 195 RPM patients from the hold-out validation set.

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

AUC in ROC, AUC in PR

**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**

Biomarkers, which were collected quarterly during the study, contained up to 11.4%missing values. All other predictors contained less than 5%missing values. For daily transmitted variables, we used forward-filling to impute missing values where previous recordings were available, and backward-filling otherwise. For variables contained in the baseline risk model, we used linear regression imputation using up to four regressors chosen based on highest correlation.

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

U

**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**

Unclear

**Rationale of bias rating**

Considerable amount of patients, however very little positive outcomes (HF hospitalisations). Exact amount of positive outcomes unclear. Good validation via hold-out set.

**Overall Risk of bias**

High