**Paper 1**

**Title: The TIMI Risk Score for Unstable Angina/Non-ST Elevation MI**

**Clinical utility**

Explanation or rationale of why these prediction models are being built or developed?

* It seems so – “multivariate analyses that adjust for several prognostic variables simultaneously provide a more accurate tool for risk stratification” and “readily applicable using standard patient features that are part of routine medical evaluation”

Was the study aim clear?

* “report the development, testing, and clinical utility of a risk stratification tool for evaluation of patients with UA/NSTEMI”

Is it prediction of outcome, or identification important features?

* Identification of impt features

**Data source and study description**

How was the data collected?

* “Using the database of the thrombolysis in myocardial infarction (TIMI) 11B trial, a phase 3 trial comparing low-molecular-weight heparin (enoxaparin) with un-fractionated heparin”

What was the study design?

* Phase 3 RCT

RCT, observational, cross-sectional, longitudinal, nationally representative survey?

* RCT

Study start, end dates reported?

* Only in abstract

What was the baseline?

* Unclear to me

Are the data measurable in clinical setting routinely or they are measured irregularly?

* Measured in “routine medical evaluation”

**Target population**

Was it clear who was the target population where this model was developed and where it can be generalized?

* Partients eligible for the trials (extensive eligibility and exclusion criteria provided) were the target population
* Significant screening means low generalizability to general population (this is acknowledged in limitations)

**Analytic data**

How was the data pre-processed?

* Personally unsure

Inclusion, exclusion criteria properly implemented to properly target the intended population?

* It seems so, but these were done as part of the two trials from which the data were generated

Clinicians were consulted to discuss the appropriateness of inclusion, exclusion criteria?

* Not reported here

Protocol published a priori?

* Don’t know.

**Data dimension, and split ratio**

Total data size, - TIMI 11B n = 3910, ESSENCE n = 3171

analytic data size, -

training,

tuning,

testing data size? – 1957 assigned to receive unfractionated hep in TIMI 11B

**Outcome label**

How was the gold standard determined, and what was the quality? The prediction of such outcome clinically relevant?

**Features**

How many covariates or features used?

* 7

How were these variables selected?

* Using multivariate logistic regression (stepwise backward elimination)
* Evaluated using C statistic (AUC from ROC)

Subject area experts consulted in selection and identification of some or all of these variables?

* unclear

Any of these variables transformed or dichotomized or categorized or combined?

* All 12 variables dichotomized – only 7 made it to multivariable logistic regression

A table of baseline characteristics of the subjects, stratified by the outcome labels presented?

* NO

**Missing data**

Were the amount of missing observations reported?

* no

Any explanation of why they were missing?

* No

How were the missing values handles?

Complete case or multiple imputation?

**ML model choice**

Rationale of the ML model choice (logistic, LASSO, CART or extensions, ensemble, or others)?

* No rationale provided, other than the fact that the authours believe multivariable analyses are a mor acurate way to stratify risk

Model specification?

Additive, linear or not?

Amount of data adequate given the model complexity (number of parameters)?

**ML model details**

Details about ML model and implementation reported? Model fine tuned? Model somehow customized? Hyperparameters provided?

**Optimism or overfitting**

What method was used to address these issues? What measures of performances were used? Was there any performance gap (between tuned model vs internal validation model)? Model performance reasonable, or unrealistic?

**Generalizability**

External validation data present? Model was tested in real-world clinical setting?

**Reproducibility**

repeatable and reproducible? These can be in 3 levels (i) model (ii) code (iii) data or their combinations.

* No

Software code provided?

* no

Which software and version was used?

* SAS PROC Logistic?

Was the computing time reported?

* no

**Interpretability**

Clinicians were consulted? Results were interpreted in collaboration with clinicians and subject area experts? Model results believable, interpretable?

**Subgroup**

Clinically important subgroups considered?

**Paper 2**

**Title: Machine Learning Prediction of Death in Critically Ill Patients With Coronavirus Disease 2019**

**Clinical utility**

Explanation or rationale of why these prediction models are being built or developed?

Was the study aim clear?

Is it prediction of outcome, or identification important features?

**Data source and study description**

How was the data collected?

* variables from the first 48 hours of ICU admission
* manual chart review using a standardized case report form. These data included demographic information, comorbidi- ties, symptoms, vital signs on ICU admission, longi- tudinal laboratory values and physiologic parameters, and outcomes. Hospital-level data included the number of pre-COVID ICU beds

What was the study design? RCT, observational, cross-sectional, longitudinal, nationally representative survey?

* observational

Study start, end dates reported?

* March 4, 2020, and June 29, 2020

What was the baseline?

Are the data measurable in clinical setting routinely or they are measured irregularly?

* Routine

**Target population**

Was it clear who was the target population where this model was developed and where it can be generalized?

* STOP-COVID is a multicenter cohort study that enrolled consecutive adult ICU patients with COVID-19 from 68 U.S. hospitals, including a variety of hospital sizes and types across a wide geographic range

**Analytic data**

How was the data pre-processed? Inclusion, exclusion criteria properly implemented to properly target the intended population? Clinicians were consulted to discuss the appropriateness of inclusion, exclusion criteria? Protocol published a priori?

**Data dimension, and split ratio**

Total data size, analytic data size, training, tuning, testing data size?

**Outcome label**

How was the gold standard determined, and what was the quality? The prediction of such outcome clinically relevant?

**Features**

How many covariates or features used? How were these variables selected? Subject area experts consulted in selection and identification of some or all of these variables? Any of these variables transformed or dichotomized or categorized or combined? A table of baseline characteristics of the subjects, stratified by the outcome labels presented?

**Missing data**

Were the amount of missing observations reported? Any explanation of why they were missing? How were the missing values handles? Complete case or multiple imputation?

**ML model choice**

Rationale of the ML model choice (logistic, LASSO, CART or extensions, ensemble, or others)? Model specification? Additive, linear or not? Amount of data adequate given the model complexity (number of parameters)?

**ML model details**

Details about ML model and implementation reported? Model fine tuned? Model somehow customized? Hyperparameters provided?

**Optimism or overfitting**

What method was used to address these issues? What measures of performances were used? Was there any performance gap (between tuned model vs internal validation model)? Model performance reasonable, or unrealistic?

**Generalizability**

External validation data present? Model was tested in real-world clinical setting?

**Reproducibility**

repeatable and reproducible? These can be in 3 levels (i) model (ii) code (iii) data or their combinations. Software code provided? Which software and version was used? Was the computing time reported?

**Interpretability**

Clinicians were consulted? Results were interpreted in collaboration with clinicians and subject area experts? Model results believable, interpretable?

**Subgroup**

Clinically important subgroups considered?