**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”
* Develop a risk score that be used by clinical staff (simple) – found in objective of abstract

Is it prediction of outcome, or identification important features?

* Identification of impt features for use in scoring tool?

**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

No internal validation in this paper – therefore no optimism correction…

External validation was present

Sample size is too small for ML methods

**Outcome label**

* **Composite of multiple outcomes**

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?

* 12, then 7 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?

* Monte-carlo simulations – very questionable in here
* If they’re not doing

Complete case or multiple imputation?

* None apparently???

**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?

* Yes in table

Model fine tuned?

* ???

Model somehow customized?

* unclear

Hyperparameters provided?

**Optimism or overfitting - if using same data to train and test model, we will see optimism of the model – we should do optimism correcting activities (sample splitting, k-sample cross validation)**

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?

* No protection against overfitting

**Generalizability**

External validation data present? (important for generalization) Model was tested in real-world clinical setting?

NOTE: If data has been used is model building, then using it to validate is internal validation

If data is new (new timeline, new source) then external.

**Reproducibility**

repeatable and reproducible? These can be in 3 levels (i) model (ii) code (iii) data or their combinations. (data access is VERY rare in medical literature)

* No
* Details about model sufficient

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?

* Yes well interpreted and compared to literature

**Subgroup**

Clinically important subgroups considered?

* No clear subgroup analysis here

NOTE: this is not essential here – important when model population is very different than

**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?

* Critically ill COVID-19 patients are overwhelming healthcare systems. Resource utilization, ICU triage, and goals of care discussions are important to these patient’s care, but the variability in these patients’ mortality complicate this process. A simple, validated, risk of death scoring system would help to improve care.

Was the study aim clear?

* Aim: Compare machine learning methods for predicting 28-day mortality in critically ill COVID-19 ICU patients, identify important risk factors for death, and develop and validate a simple validated tool for predicting 28-day mortality that could be calculated at the bedside

Is it prediction of outcome, or identification important features?

* Prediction of outcome (death within 28 days) AND identification of important features for risk score development

**Data source and study description**

How was the data collected?

* From the cohort of participants of the STOP-COVID multicenter cohort study that enrols 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
* Data extracted using manual chart review using a standardized case report form. Data included demographic information, comorbidities, symptoms, vital signs on ICU admission, longitudinal laboratory values and physiologic parameters, and outcomes. Hospital-level data included the number of pre-COVID ICU beds.
* variables from the first 48 hours of ICU admission

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

* Observational study (COHORT study participant chart data)

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?

* The target population is clear – critically ill adult patients with COVID-19

**Analytic data**

How was the data pre-processed? (see section before splitting into training and testing)

* Predictor variables include those extracted from patient charts that were collected within the first 2 days of their admission to the ICU. Predictor variables were excluded from consideration in the model immediately based upon their presence (missing data considerations) – if a variable was more than 50% missing it was not considered (ex: interleukin, fibrinogen), if a variable had rare positivity (i.e. <5% of individuals had it) it was not considered to be part of the model (e.g. the use of mechanical cardiac support devices)
* Outcome variable (inhospital death within 28 days of ICU admission) was reported. Note that patients discharged alive before day 28 were assumed to be alive at day 28. Validity of this assumption confirmed

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

* Yes, inclusion criteria (Adult (>18) and with laboratory-confirmed COVID-19 admitted to an ICU at a participating site between March 4, 2020, and June 29, 2020, were eligible for inclusion) was reported. Exclusion criteria not stated but implied. The included population properly targets intended population.

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

* Not clear, but likely given the authors’ expertise and training.

Protocol published a priori?

* Not clear

**Data dimension, and split ratio**

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

* Total data size - 5,075 patients included in the study, 1,846 (36.4%) of whom died by day 28. Recruited from 68 ICU’s across the USA
* **Randomized split** of 68 ICU’s into two groups (one containing 75% of ICU’s and one containing 25% of ICU’s – proportions arbitrary/unjustified) – 75% is the “training data”, and 25% is the data for externally validating.
* Individuals recruited between March and April 2020 were considered for the primary “external” validation (as they are completely different sites)
* For the primary external validation analysis, 51 sites (*n* = 3,825 admission) were included as training data and 17 sites were included for independent validation (*n* = 810 admissions) using data from March 2020 to April 2020
* Secondary temporal validation, which used the same models from the training dataset above (*n* = 3,825 admission), validated the models in 440 admissions from May 2020 to June 2020

Notes from lecture:

* Bit confusing here – 75 25 split is not the same as sample splitting (KEY POINT)
* Very common to split into training and testing (typically, just pool and then split)
* Not clear why they did 75/25 and not clear why they did this at level of the hospital
* Tuning (hyperparameters?) done based on only training data

**Outcome label**

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

* Binary outcome = dead or not
* Assume to be alive if they were discharged before 28 days
* Hard outcome (easy to get in ICU)
* Definitely clinically relevant – knowing information about an individuals likelihood of heath can help direct care resources to prevent such event

**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?

* Unclear whether covid-19 experts were involved in identification of variables
* Features (aka predictor variables) included are those collected in the first 48 hours of ICU admission for a typical patient – there was some exclusion of variables based on data missingness (large amounts) and low positivity (as mentioned above).
* Final list of covariates considered includes: age, vital signs and respiratory support on ICU admission, Fio2 among patients requiring invasive mechanical ventilation, laboratory values, and organ support.
* A table of baseline characteristics is in the appendix (eTable 6) detailing subjects characteristics stratified by outcome

**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?

* There was missing data – reported well in appendix
* Missing data for ML model predictors were imputed using bagged trees method (using mode)
* Big differences in missing data between training data, internal test data, and external test data
* For simple clinical tool development, Missing values were imputed using the mode of each variable category in the training data to make it easier to operationalize at the bedside. NOTE: its very unclear why they needed to use lasso regression to select most important variables for this simple clinical tool, when XGboost model already identified and ranked important variables (could have just took top ten instead)

**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)?

* Model XG boost choosen because it had the highest AUC when externally validated (both ways – using different data (the 25%) and temporally)). This means it wascapable of distinguishing between death positives and negatives best.

**ML model details**

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

* 10 fold cross validation was used to select hyperparameters ONLY USING TRAINING DATA – unusual?
* Whole data cross validation would have been a more common approach
* Potential for bias or error because of only using training data
* Same discussion for temporal validations

**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?

* Authours utilized 10 fold cross-validation (internal validation) to maximize AUC, but only in the training data
* Authours also utilized the 25% of hospitals/data not used to train to externally validate models (where AUC was highest for XG boost)
* Authours also completed temporal validation (in data from the 440 admissions occurring between May and June 2020)

**Generalizability**

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

* External validation used 25% of whole data that was not used in training
* Temporal validation too.

**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?

* No code
* No data
* Model parameter details are provided in depth in eTable 8
* With data request we may be able to reproduce
* Software used is indicated (R), as well as package used (caret) and the models within such package
* Computing time was NOT reported

**Interpretability**

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

* Reasonably interpretable, simple model that could use a simple calculator – very likely to be interpretable
* Little discussion of clinical experts involved in interpretation, however, most authors are medically trained and would likely have contributed to interpretation.

**Subgroup**

Clinically important subgroups considered?

* Subgroups are not considered/relevant? here

**MEDI 504A – Lab 5 – Critical Appraisal of Data Science in Health studies**

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

**By**: Matthew Manson

**Paper Summary:** Patients who are critically ill with COVID-19 are observed to have variable mortality. Understanding a patient’s risk of death using variables from the first two days of Intensive Care Unit admission would enable care teams to distribute care resources and facilitate appropriate care trajectory discussions more appropriately. Author’s Churpek et al. sought to compare machine learning algorithms for predicting death and utilize such algorithms to identify important predictor variables that could be incorporated into a simple mortality prediction tool for use at the bedside. To do so, they utilized chart data from adult patients with lab confirmed COVID-19 who received care between March 4, 2020, and June 29, 2020 in one of the 68 U.S. ICU’s participating in the STOP-COVID cohort study. 5075 patients were included in the analysis, which revealed eXtreme Gradient Boosting to have the highest AUC in both external and temporal validation. Important variables, including age, number of ICU beds, creatinine, lactate, arterial pH, and Pao2/Fio2 ratio, were incorporated into a simple tool for use at bedside that predicted death better than the Sequential Organ Failure Assessment score, National Early Warning Score, and CURB-65. The author’s simple tool could be used to improve triage decisions and provide prognostic information for both patients and care providers.

**Critical Appraisal following Dr. Ehsan Karim’s Key Considerations when appraising a Machine Learning Research Article:**

Author’s Churpek et al. provide a generally well done account of the work they completed to develop the simple bedside death prediction scoring system for critically ill COVID-19 patients. Below is a point-by-point commentary on each significant consideration of a well reported ML research article.