External validation of prognostic models in critical care: a cautionary tale from covid-19 pneumonitis

Sebastian Bate, Victoria Stokes, Hannah Greenlee, Kwee Yen Goh, Graham Whiting, Gareth Kitchen, Glen P Martin, Alexander J Parker, and Anthony Wilson, on behalf of the MRI Critical Care Data Group

Manchester University NHS Foundation Trust

Background

Back in the deep dark days of March 2021, we discussed making a new prognostic model for ICU. It was the highlighted that we should actually look to see if there were any models that actually worked.

We started with looking at the UCL model (Arina, 2020) which used CRP and BNP in a model for CPAP failure (death/intubation) vs success. External validation in Manchester using two cohorts: patients up to 15/6/20 (when dexamethasone was introduced as routine) and all patients to 31/10/21. Model works well among patients pre-dexamethesone (C=0.839) but poorly after (C=0.613) (Stokes, 2022).

We then asked the question, would we find the same results in other models (pre-dex good, post-dex bad)?

Aims

1. Are published ICU multivariable prediction models sufficiently described to allow new predictions to be generated, and performance compared to the original publication?
2. Do these models use parameters that are routinely recorded in most ICUs?
3. Is the performance (by discrimination and/or calibration) of these models replicable in patients from a similar period to when the model was developed, either pre- or post-the introduction of dexamethasone?
4. Is the performance (by discrimination and/or calibration) of these models replicable using patients for the duration of the pandemic?

Methods

Lit Search: Wynants (2020), PubMed, reverse citation

IC/EC: models where the study population used to develop the original model was

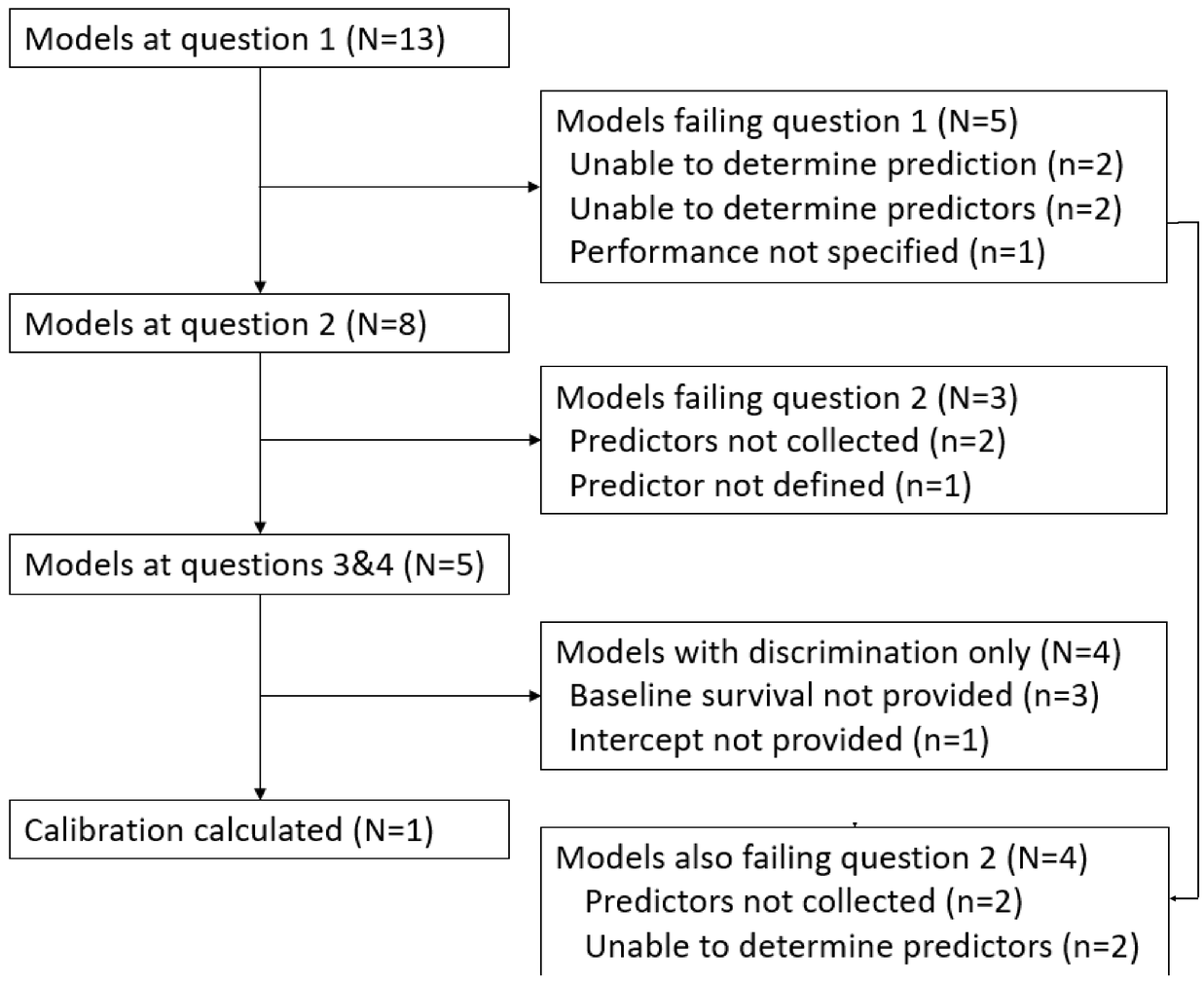
- Adults (aged over 16 years) admitted to the ICU with a diagnosis of COVID pneumonitis
- Required at least two predictor variables, and which were measured within 24 hours of, or prior to, ICU admission or 24 hours prior to onset of invasive mechanical ventilation (IMV), depending on the model, or demographic data.
- The endpoint was death at any point, need for IMV, or length of stay in ICU.

Data extraction: assess whether predictions can be generated based off what was published; cross reference variables against what was available in the EPR

Those that qualify: remake the model using the IC/EC of the original publication; check discrimination using the same test that the original paper used and calibration (if possible).

Data: Manchester Royal Infirmary ICU EPR (1/3/20 to 28/2/22); incidental covid and national data opt-out excluded; ethical approval 21/HRA/3518

Results



	Endpoint	Predictors used	Development cohort size	Model Type	Original publication		Pre-dexamethasone era		All patients	
Reference					N	Reported C	N	C	N	C
Arina	IMV or death	CRP, BNP	93	Logistic regression	93	0.804 (0.728 - 0.880)	27	0.793 (0.618-0.968)	103	0.596 (0.482-0.710)
Cao	Death	Urea, hs-CRP	77	Logistic regression	77	0.857 (0.77–0.94)	22	0.567 (0.321-0.813)	141	0.558 (0.460-0.655)
Leoni, Lombardelli	Death censored at 28 days	Age, Obesity, Procalcitonin, SOFA, PaO_2/FiO_2	229	Cox regression	229	0.821 (0.766–0.876)	33	0.605 (0.471-0.739)	204	0.564 (0.506-0.622)
Leoni, Moschini	Death censored 28 days	Mnutric, hs-CRP, Neutrophils	98	Cox regression	98	0.720 (0.67-0.79)	34	0.560 (0.428-0.692)	201	0.546 (0.484-0.608)
Moisa	Death censored at 28 days	Age, NLR, SOFA	425	Cox regression	425	0.697 (0.755–0.833)	53	0.672 (0.574-0.770)	291	0.686 (0.640-0.732)

BNP, N-terminal pro B-type natriuretic peptide; C, concordance statistic (hs-)CRP, (high sensitivity)y C-reactive protein; IMV, invasive mechanical ventilation; Mnutric, modified nutritional risk in critically ill; NLR, neutrophil-to-lymphocyte ratio; SOFA, serial organ failure assessment.

Too many models can't be validated, of those only one retained performance but still cannot be recommended for clinical use.

Conclusion

- In ICU, where the pressure on beds is high and the decisions around such resource use so important, clinical prediction models are particularly appealing and yet demonstrably unhelpful at present.
- On the basis of our findings, we urge bedside intensivists to exercise caution before using any of the examined models to support clinical decision making.
- Don't go round making prediction models because you are bored
- Use TRIPOD to at least give us a fighting chance of being able to validate it

References and supplementary info

/sebastianbate @sebastianbate sgbstats