A calibration hierarchy for risk models: strong calibration occurs only in utopia

Ben VAN CALSTER[a,b], Daan NIEBOER[b], Yvonne VERGOUWE[b], Bavo DE COCK[a], Michael J. PENCINA[c,d], Ewout W. STEYERBERG[b]

[a] KU Leuven, Department of Development and Regeneration, Leuven, Belgium
[b] Department of Public Health, Erasmus MC, Rotterdam, the Netherlands
[c] Duke Clinical Research Institute, Duke University, Durham (NC), USA
[d] Department of Biostatistics and Bioinformatics, Duke University, Durham (NC), USA

APPENDIX

Appendix 1

Moderate calibration guaranties non-harmful decisions in terms of Net Benefit and decision curve analysis

Notation:

$\pi(\mathbf{x})$: true probability of event given covariate vector **x**, where $\mathbf{x} = (x_1, \ldots, x_i)$

$\hat{\pi}(\mathbf{x})$: estimated probability of event given covariate vector **x** given by a prediction model.

Y: outcome (1=event, 0=non-event)

$f(\mathbf{x})$: probability density of covariate vector **x**

P_1 and P_0: event rate $P(Y = 1)$ and 1 minus event rate $P(Y = 0)$

For a moderately calibrated model the following property holds, among patients with an estimated risk of R% on average R out of 100 of these patients have the event.

Mathematically this can be translated as:

$$E[Y|\hat{\pi}(\mathbf{x}) = r] \quad = \frac{\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r} \pi(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}}{\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r} f(\mathbf{x}) d\mathbf{x}} = r$$

The expected net benefit at threshold t is given by:

$$E[\mathrm{NB}_t] \quad = E\left[\frac{N_{TP} - \mathrm{odds}(t) N_{FP}}{N}\right] = P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - \mathrm{odds}(t) P(Y = 0, \hat{\pi}(\mathbf{x}) \geq t).$$

To ensure that a moderately calibrated model is not harmful, the expected net benefit should be larger than (a) the net benefit of treating all patients and (b) the net benefit of treating no patients.

$$P(Y=1,\hat{\pi}(\mathbf{x})\geq t)=\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}\pi(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

$$=\int_{t}^{1}\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r}\pi(\mathbf{x})f(\mathbf{x})d\mathbf{x}dr$$

$$=\int_{t}^{1}r\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r}f(\mathbf{x})d\mathbf{x}dr$$

$$\geq t\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}f(\mathbf{x})d\mathbf{x}.$$

$$P(Y=0,\hat{\pi}(\mathbf{x})\geq t)=\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}(1-\pi(\mathbf{x}))f(\mathbf{x})d\mathbf{x}$$

$$=\int_{t}^{1}\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r}(1-\pi(\mathbf{x}))f(\mathbf{x})d\mathbf{x}dr$$

$$=\int_{t}^{1}(1-r)\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})=r}f(\mathbf{x})d\mathbf{x}dr$$

$$\leq(1-t)\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}f(\mathbf{x})d\mathbf{x}.$$

From above we get:

$$\begin{aligned}E[\mathrm{NB}_t] &= P(Y=1,\hat{\pi}(\mathbf{x})\geq t)-\mathrm{odds}(t)P(Y=0,\hat{\pi}(\mathbf{x})<t)\\ &\geq t\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}f(\mathbf{x})d\mathbf{x}-\frac{t}{1-t}(1-t)\int_{\mathbf{x}:\hat{\pi}(\mathbf{x})\geq t}f(\mathbf{x})d\mathbf{x}=0\\ &=0\end{aligned}$$

So with moderate calibration using the model is better than treating no patients. Remains to show that for t below the event rate (P_1) using the model gives a higher expected net benefit than treating all patients.

$$E[NB_t] - E\left[NB_{\text{treat all},t}\right] = P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - \frac{t}{1-t} P(Y = 0, \hat{\pi}(\mathbf{x}) \geq t) - \left(P_1 - \frac{t}{1-t} P_0\right)$$
$$= (P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - P_1) - \frac{t}{1-t}(P(Y = 0, \hat{\pi}(\mathbf{x}) \geq t) - P_0)$$

To achieve a better NB using the model compared to treating all patients we then need

$$(P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - P_1) \geq \frac{t}{1-t}(P(Y = 0, \hat{\pi}(\mathbf{x}) \geq t) - P_0)$$

whenever $t < P_1$.

$$\frac{t}{1-t}(P(Y = 0, \hat{\pi}(\mathbf{x}) \geq t) - P_1) \leq \frac{t}{1-t}\left[(1-t)\int_{\mathbf{x}:\hat{\pi}(\mathbf{x}) \geq t} f(\mathbf{x}) d\mathbf{x} - (1 - P_1)\right]$$
$$= t\int_{\mathbf{x}:\hat{\pi}(\mathbf{x}) \geq t} f(\mathbf{x}) d\mathbf{x} - \frac{(1-P_1)}{1-t} t$$
$$\leq P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - P_1 \frac{(1-P_1)}{1-t}$$
$$\leq P(Y = 1, \hat{\pi}(\mathbf{x}) \geq t) - P_1$$

Figure A1. Flexible calibration curves to simulate external validation of a strongly calibrated model in 50 randomly drawn datasets of size 100 to 1000 (true event rate 50%).

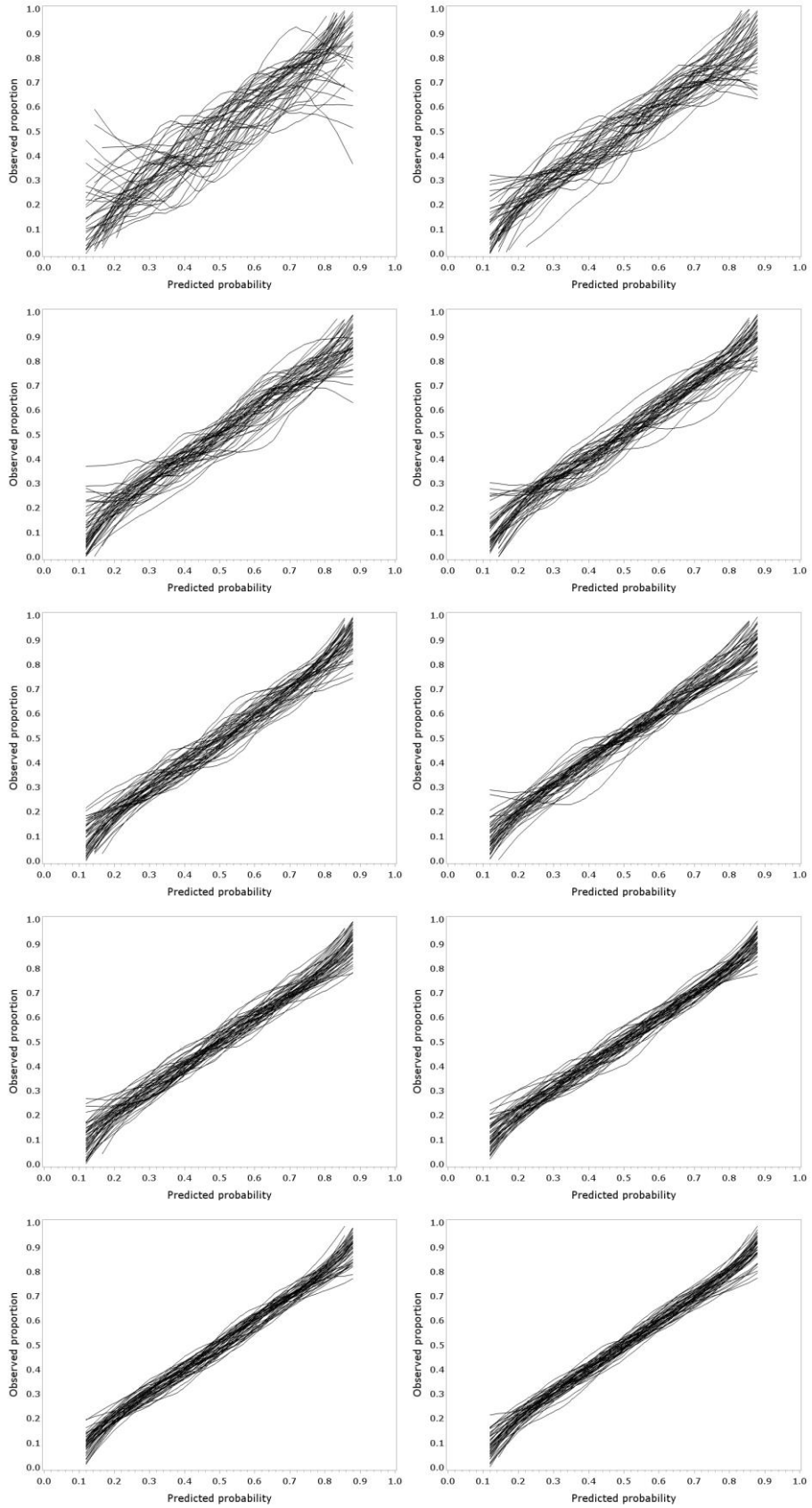
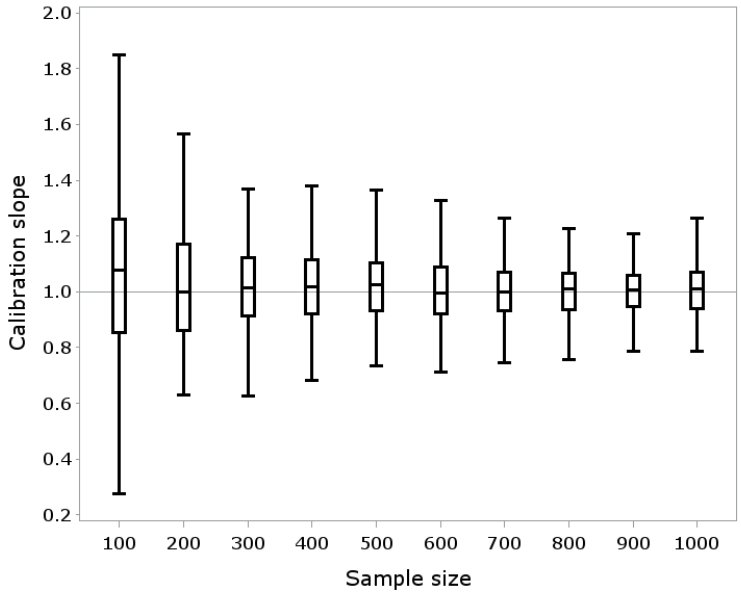


Figure A2. Box plots of the calibration slope (A) and estimated calibration index (ECI) (B) to simulate external validation of a strongly calibrated model in in 200 randomly drawn datasets of size 100 to 1000 (true event rate 50%).

(A)



(B)