Using routine blood test results to predict the risk of death for emergency medical admissions to hospital: an external model validation study

Muhammad Faisal[1,5], Robin Howes[2], Ewout W. Steyerberg[3], Donald Richardson[4], Mohammed A Mohammed[1,5,6,*]

[1]Faculty of Health Studies, University of Bradford, Bradford, UK

[2]Northern Lincolnshire and Goole NHS Foundation, Diana, Princess of Wales Hospital Grimsby, North East Lincolnshire, UK

[3]Dept of Public Health, Erasmus University, Rotterdam, The Netherlands

[4]MBChB, Department of Renal Medicine, York District Hospital, York , UK

[5]Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK

[6]Yorkshire and Humberside Academic Health Sciences Network, UK.

Muhammad Faisal M.Faisal1@bradford.ac.uk

Robin Howes robin.howes@nhs.net

Ewout W. Steyerberg E.Steyerberg@ErasmusMC.nl

Donald Richardson drichardson@doctors.org.uk

Mohammed A Mohammed M.A.Mohammed5@bradford.ac.uk

Funding: There is no specific funding for this research.

Conflict of Interest: The authors declare that there are no conflicts of interest.

*Crossponding Author

Abstract

Background

The Biochemistry and Haematology Outcome Model (BHOM) relies on the results from routine index blood tests to predict the patient risk of death. We aimed to externally validate the BHOM model.

Method

We considered all emergency adult medical patients who were discharged from Northern Lincolnshire and Goole (NLAG) hospital in 2014. We compared patient characteristics between NLAG (the validation sample) and the hospital where BHOM was developed. We evaluated the predictive performance, according to discriminative ability (with a concordance statistic, c), and calibration (agreement between observed and predicted risk).

Result

There were 29 834 emergency discharges of which 24 696 (83%) had complete data. In comparison with the development sample, the NLAG sample was similar in age, blood test results, but experienced a lower mortality (4.7% vs 8.7%). When applied to NLAG, the BHOM model had good discrimination (*c*-statistic 0.83 [95% CI 0.823 - 0.842]). Calibration was good overall, although the BHOM model overpredicted for lowest (<5%, observed = 229,predicted =286) and highest (≥50%, observed = 31, predicted = 49) risk groups, even after recalibrating for the differences in baseline risk of death.

Conclusion

Differences in patient case-mix profile and baseline risk of death need to be considered before the BHOM model can be used in another hospital. After re-calibrating for the baseline difference in risk the BHOM model had good discrimination but less adequate calibration.

Keywords: Critical Care; Emergency Medicine; Biochemistry & metabolism

Introduction

Statistically derived risk equations are widely used to support healthcare professionals in the research, audit and delivery of healthcare. Examples of risk equations include Acute Physiology and Chronic Health Evaluation (APACHE),[1] Mortality Probability Model (MPM),[2] and for more examples see.[3] Typically the risk equations are developed by randomly splitting the data into two parts – "training" and "testing". This approach, known as internal validation,[4] has been criticised because (a) the subsequent model performance statistics are optimistic and (b) typically, the use of the risk equation is beyond the settings where the equation was first developed and internal validation does not give any indication about the performance of the model in another setting. The use of external validation, where the model is tested using data from another setting, is now seen as an important step in model development. [4–8]

The Biochemistry and Haematology Outcome Model (BHOM) was developed by researchers at Portsmouth Hospital NHS Trust based on routinely collected biochemistry and haematology blood test results along with basic demographic information for 9 497 adults discharged from a medical speciality hospital during one year (January 2001 – December 2001). A major advantage of the BHOM model is that the covariates are clinically meaningful, collected as part of the process of care and these data are available within a few hours of the patient admission. The BHOM model was internally validated and found to have good discrimination (*c*-statistics 0.757 to 0.779) and good calibration (non-significant p-values from the Hosemer-Lemeshow deciles of risk table) [9 -10]. Nonetheless the BHOM model has not been externally validated – an important step before it can used outside of the hospital in which it was developed.

We aimed to externally validate the BHOM model, by considering its calibration and discrimination in a cohort of patients discharged from another hospital following an acute admission.

Methods

Setting & data for external validation

Our cohort of emergency admissions is from the Northern Lincolnshire and Goole NHS Foundation Trust (NLAG) in England. All 24 696 adults (age≥16 years) emergency patients discharged during the year 2014 (1 January 2014 to 31 December 2014) were included. For each admission we obtained the patients age, gender (male/female), discharge status (alive/dead) and index blood test results used in the BHOM model from the hospital computer system. The covariates set was:- *age* on admission (years), *gender* (female=0/male=1), *albumin* (g/L), *creatinine* (µmol/L), *haemoglobin, potassium* (mmol/L), *sodium* (mmol/L), *white cell count* (10^9 cells/L), *urea* (mmol/L), and ratio of *urea* (mmol/L) and *creatinine* (µmol/L). We also considered records which had no missing data (24 696/29 834 (83%)), as did here.[9] We did not consider elective patients because the intended use of the BHOM model in NLAG was for acute medical patients.

The BHOM Model

The BHOM risk equation is shown below:-

$$log\left(\frac{R}{1-R}\right) = -10.192 - (0.013 \times gender) + (5.712 \times mode\ of\ admission) + (0.053 \times age\ on\ admission) + (0.018 \times urea) - (0.001 \times sodium) - (0.101 \times potassium) - (0.047 \times albumin) - (0.037 \times haemoglobin) + (0.067 \times white\ cell\ count) + (0.001 \times creatinine) + (2.744 \times urea/creatinine)$$

Where *R* is risk of death in hospital and the variables *gender* and *mode of admission* are coded female=0, male=1, elective=0, and emergency=1, respectively. The other covariates are continuous values based on routine blood test results.

Statistical analyses

We followed a previously proposed framework for the external validation of clinical prediction models.[6] There are two key steps:

1. To determine the relatedness of the patients in the model development sample with the patients in the external validation sample. This preliminary step helps to determine the extent to which the model is being validated in a patient population that is not materially dissimilar to the development sample.
2. To assess the performance of the model on the external validation sample by determining the model discrimination and calibration. For *discrimination*, we use the area under the receiver-operator curve (AUROC) or concordance (c)-statistic. AUROC is the probability that the model will predict a higher risk of death for a randomly selected patient who died, compared to a randomly selected patient who survived.[4-5] *Calibration* is the relationship between the observed and predicted risk of death and can be usefully seen on a scatter plot (y-axis observed risk, x-axis predicted risk). Perfect predictions should be on the 45° line. The intercept (a) and slope (b) of this line gives an assessement of 'calibration-in-the-large'.[11] At model development, $\mathrm{a} = 0$ and $\mathrm{b} = 1$, but at validation, calibration-in-the-large problems are indicated if a is not 0 and if b is more/less than 1 as this reflects problems of under/over prediction.[5] We also use the Hosmer–Lemeshow (HL) goodness of fit test for calibration with degree of freedom $df = g - 1$,[4] where g is deciles of risk groups as defined by Prytherch et al., 2005 (see Table 2).

Before we could apply the BHOM model we made two adjustments based on preliminary observations. (1) We divided the NLAG *haemoglobin* results by 10 to ensure they were compatible with units for *haemoglobin* in the BHOM model. (2) We noted that the mortality in NLAG is almost half that of Portsmouth Hospital (4.69% vs 8.7%). To correct for this difference in baseline risk we adjusted the constant term in the BHOM model, by trial and error (see supporting Microsoft Excel

file) and selecting the value (-11.3235) which produced optimal calibration. The mean risk of death for NLAG from the model with the revised constant was thus similar with BHOM model 4.69% [12].

Ethical Approval

The study does not require ethical approval because it meets the exemption criteria ("Research limited to secondary use of information previously collected in the course of normal care (without an intention to use it for research at the time of collection), provided that the patients or service users are not identifiable to the research team in carrying out the research.[13])"

Results

Cohort description

There were 29 834 emergency discharges from during the year 2014, of which 24 696 (83%) had complete data. The mean age of patients was 63.1 years (SD 21.1), the female to male ratio was 1.14 and the in-hospital mortality was 4.69% (1159/24696). The relationship between the continuous covariates and mortality are shown in Figure 1 and Figure 2.

External Validation Results

Step 1: Relatedness of the patient samples

The mean and SD for each continuous covariates showed no major differences between the development sample (Portsmouth Hospital) and external validation sample (NLAG), with the exception of albumin which appear to be higher in Portsmouth.

Step 2: Assessing the model performance

We applied the BHOM model to the validation sample at NLAG. The resulting *c*-statistic (discrimination) was 0.833 (95% confidence interval 0.823 to 0.843) and the calibration in-the-large

was $a = 0$ and $b = 0.99$. Calibration plots (Figure 3), without and with re-calibration for differences in baseline risk showed systematic over prediction which still persisted in the higher risk (risk >0.40) groups.

The H-L deciles of risk calibration test (see table 2), after recalibrating for baseline differences, was statistically significant p<0.001 (χ^2 = 51.49, 8 df). Over prediction was evident in the lowest risk group (≥0% to <5%, n=16804, χ^2 = 10.13, 1 df), where there were 229 observed deaths compared with 286 predicted deaths, and in the highest risk group (≥50% to <100%, n=66, χ^2 = 27.40, 1 df), where there were 31 observed deaths compared with 43 predicted deaths. Under prediction was seen in the fifth risk group (≥12.5% to <15%, n=695, χ^2 = 7.15, 1 df) where there were 119 observed deaths compared with 95 predicted deaths.

Discussion

The performance of the BHOM model based on internal validation was good - the discrimination (*c*-statistic) for BHOM was 0.757 to 0.779 and the Hosmer-Lemeshow deciles of risk calibration produced no statistically significant difference between predicted and observed mortality (p>0.05). We undertook an external validation exercise for the BHOM model using data for emergency medical admissions at NLAG hospital over one year. As far as we are aware, this is the first external validation attempt of the BHOM model. We found that after re-calibrating for the baseline difference in risk between the two cohorts of patients, the BHOM model had good discrimination, but less adequate calibration - it over predicted for lowest (<5%, observed = 229, predicted =286) and highest (≥50%, observed = 31 , predicted = 49) risk groups.

Whilst the BHOM model has attractive features of using results from routine blood tests (without additional data collection) its use outside of hospital in which it was developed requires attention to

two key issues. (1) Consideration and, where necessary, correction for differences in baseline risk by adjusting the constant term in the BHOM model. (2) Investigation of predicted versus observed risk as seen in a calibration plots, which in our case, showed that differences persisted even after correcting for baseline differences in risk of death. The inadequate calibration is not readily explained by difference in the distribution of continuous and categorical covariates. Further work to consider reasons for inadequate calibration are required. There are several possible issues. (1) The sample sizes at NLAG are almost three times as large as Portsmouth hospital (24 696 vs 9 497). This would increase the risk of spuriously low p-values which are statistically significant but clinically insignificant. (2). The calibration deteriorates in higher (≥50%) risk groups and so the model predictions could be capped at this threshold. (3) The relationship between covariates and risk of death may be significantly different in NLAG versus a Portsmouth hospital. This could be explored using tests for interactions. Nonetheless it is worth emphasising that whilst these desktop approaches are useful and can correct for some issues in model performance, the ultimate question is to determine the extent to which such risk equations support clinical decision making and enhance safety and quality of patient care.

Conclusion

Differences in patient case-mix profile and baseline risk of death need to be considered before the BHOM model can be used in another hospital. We found that after re-calibrating for the baseline difference in risk between the two cohorts of patients, the BHOM model had good discrimination, but less adequate calibration.

References

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med [Internet]. 1985 Oct [cited 2015 Feb 19];13(10):818–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3928249

2. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. JAMA. 1993;270(20):2478–86.

3. Vincent J-L, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care. 2010;14(2):207.

4. Steyerberg EW. Clinical Prediction Models. A practical approach to development, validation and updating. Springer; 2008.

5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology [Internet]. 2010 Jan [cited 2014 Jul 12];21(1):128–38. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3575184&tool=pmcentrez&rendertype=abstract

6. Debray TP a., Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol [Internet]. Elsevier Inc; 2014;68(3):279–89. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0895435614002753

7. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol [Internet]. 2014 Jan [cited 2015 May 2];14(1):40. Available from: http://www.biomedcentral.com/1471-2288/14/40

8. van de Laar R, IntHout J, Van Gorp T, Verdonschot S, van Altena a M, Gerestein CG, et al. External validation of three prognostic models for overall survival in patients with advanced-stage epithelial ovarian cancer. Br J Cancer [Internet]. Nature Publishing Group; 2014;110(1):42–8. Available from: http://www.nature.com/doifinder/10.1038/bjc.2013.717

9. Prytherch DR, Sirl JS, Schmidt P, Featherstone PI, Weaver PC, Smith GB. The use of routine laboratory data to predict in-hospital death in medical admissions. Resuscitation. 2005;66(2):203–7.

10. Badriyah T of P. Developing Risk of Mortality and Early Warning Score Models using Routinely Collected Data [Internet]. University of Portsmouth; 2013. Available from: http://eprints.port.ac.uk/13999/1/Thesis_Tessy_Badriyah_2013.PDF

11. Cox DR. Two further applications of a model for binary regression. Biometrika [Internet]. 1958 Dec 1 [cited 2015 Oct 25];45(3-4):562–5. Available from: http://biomet.oxfordjournals.org/content/45/3-4/562.citation

12. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol [Internet]. 2016 Jan 6 [cited 2016 Apr 15]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/26772608

13. NHS Health Research Authority. Governance Arrangements for Research Ethics Committees [Internet]. [cited 2015 Nov 20]. Available from: http://www.hra.nhs.uk/resources/research-legislation-and-governance/governance-arrangements-for-research-ethics-committees/

	Portsmouth Hosptial		NLAG	
Variables	**Mean**	**SD**	**Mean**	**SD**
Age (years)	63.3	20.8	63.1	21.1
Albumin (g/L)	39.7	5.7	34.0	6.2
Creatinine (µmol/L)	114.3	80.5	100.3	77.9
Haemoglobin	13.5	2.2	12.9	2.2
Potassium (mmol/L)	4.3	0.6	4.1	0.6
Sodium (mmol/L)	137.8	4.4	137.0	4.8
White cell count (10^9 cells/L)	10.4	4.9	9.9	6.9
Urea (mmol/L)	8.0	6.7	7.3	5.8
Urea (mmol/L)/ Creatinine (µmol/L)	0.07	0.03	0.07	0.03

Table 1: Relatedness of continuous covariates in BHOM and NLAG data sets

Risk group(%)	No. of cases	Mean predicted risk (%)	Predicted	Observed	χ^2
≥0 to <5	16804	1.70	286	229	10.13
≥5 to <7.5	3023	6.16	186	209	2.25
≥7.5 to <10	1801	8.65	156	175	3.14
≥10 to <12.5	1090	11.13	121	125	0.07
≥12.5 to <15	695	13.64	95	119	7.15
≥15 to <20	651	17.17	112	113	0.01
≥20 to <25	292	22.24	65	69	0.73
≥25 to <33	173	28.17	49	47	0.40
≥33 to <50	101	39.31	40	42	0.22
≥50 to <100	66	74.90	49	31	27.40
≥0 to <100	**24696**	**-**	**1159**	**1159**	**51.49**

Table 2: Hosmer-Lemeshow deciles of risk table

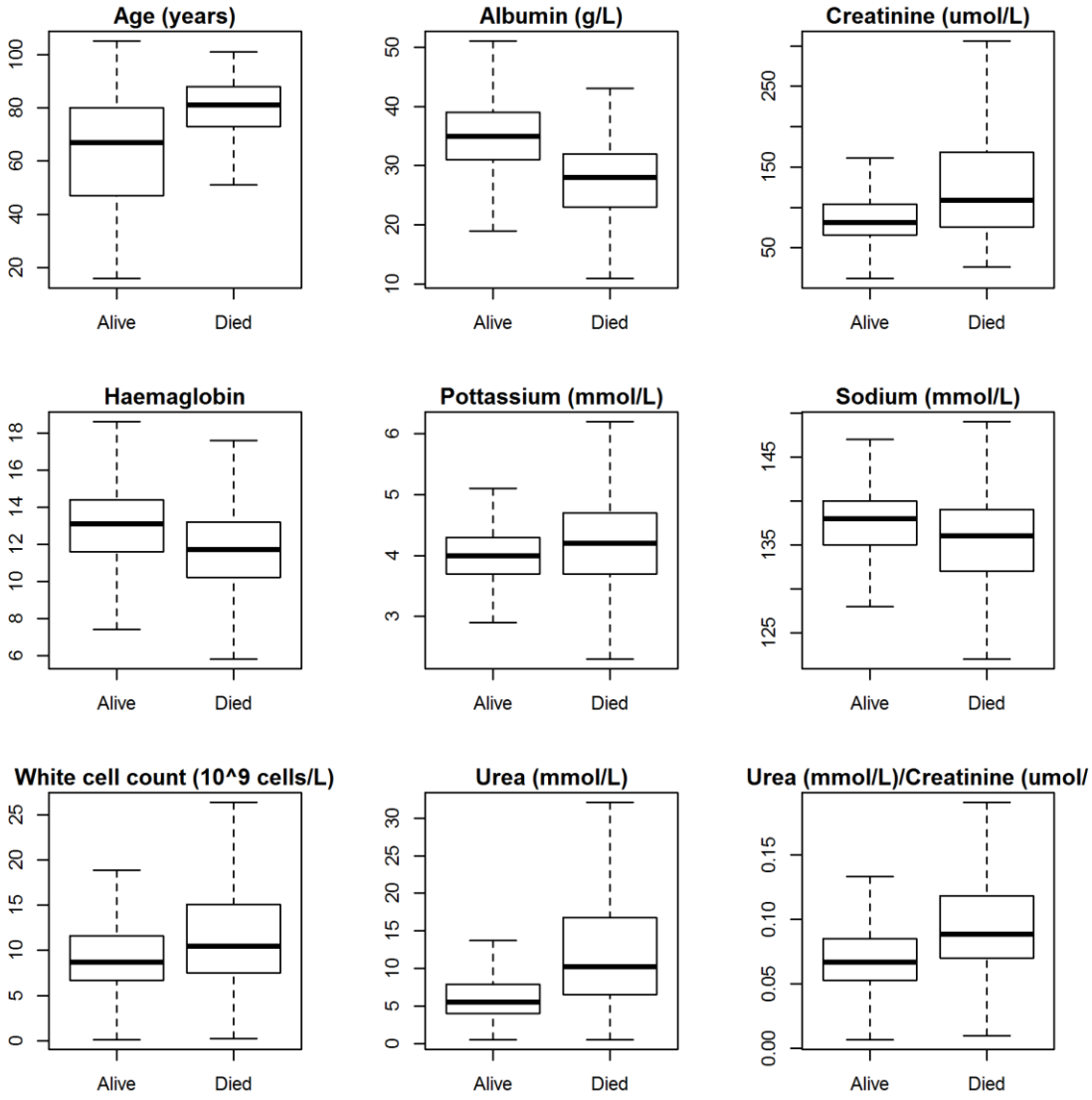

Figure 1: Boxplot without outliers for continuous covariates with respect to patient's discharge status (Alive/Died)

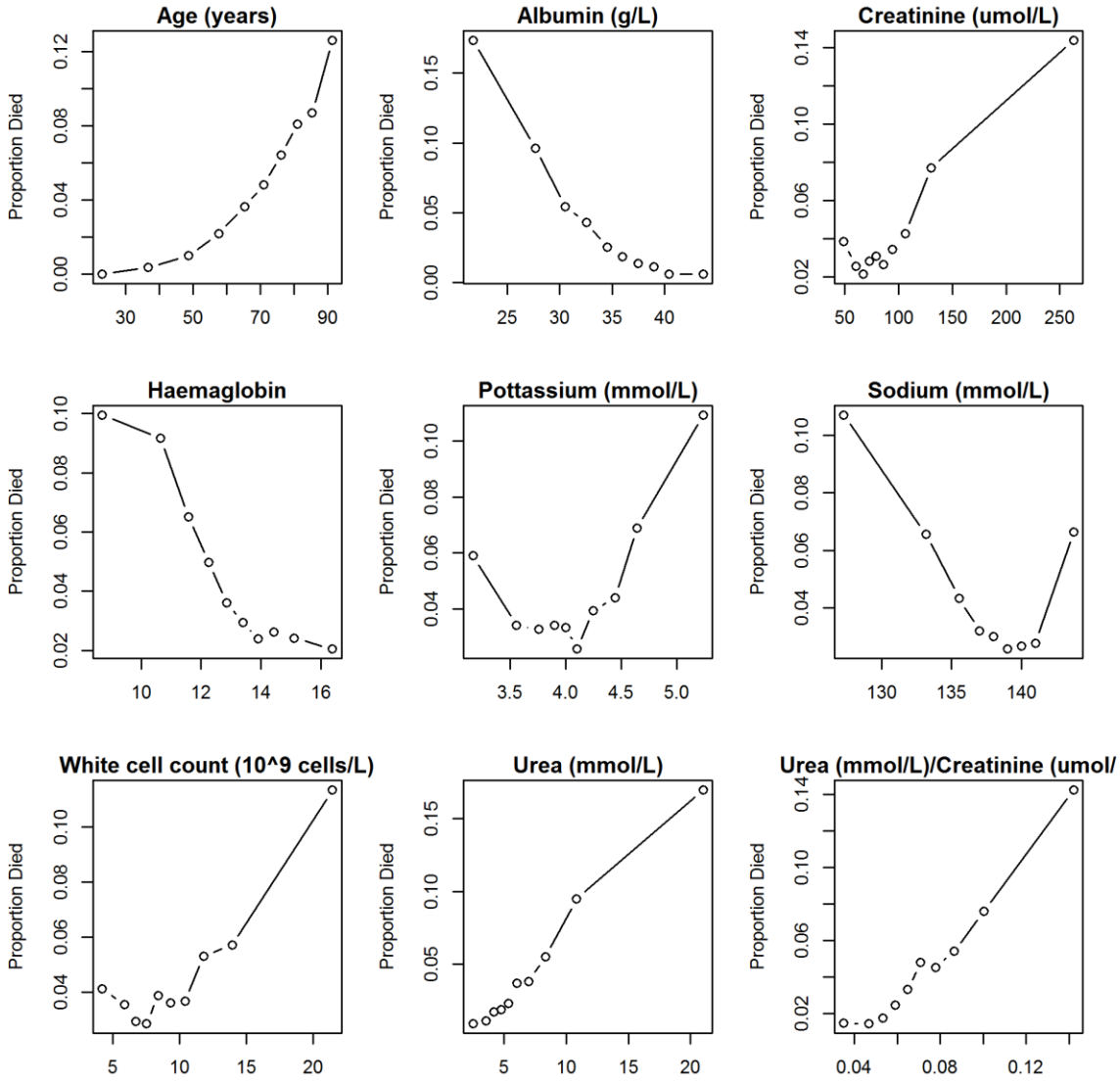

Figure 2: Scatter plots showing the observed risk of death with continuous covariates.
NB: y-axis range changes in each plot.

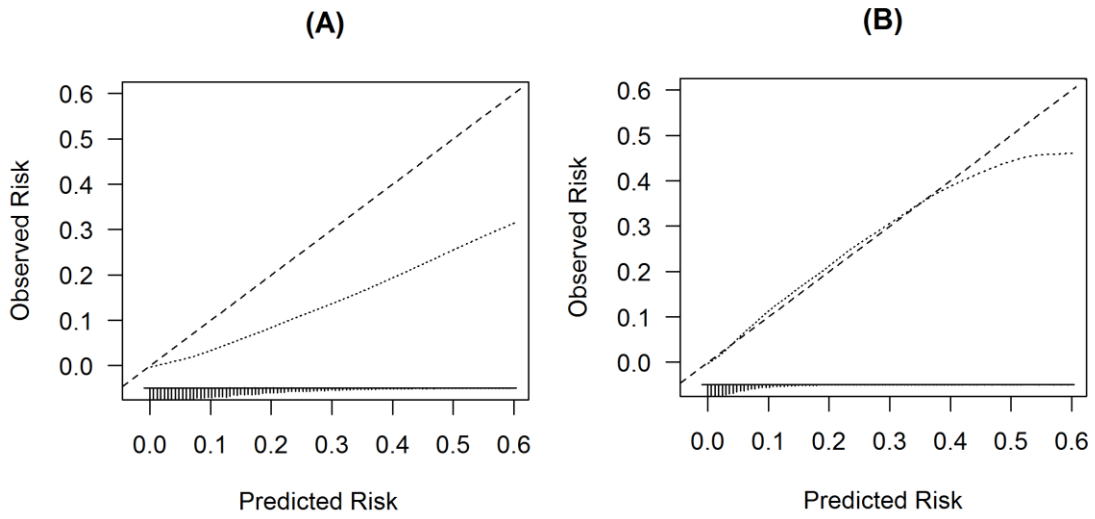

Figure 3: Calibration plots: (A) before recalibrating the BHOM model (B) after recalibrating for differences in baseline risk.
Dashed line is the ideal. Dotted line is actual.

Gender	Mode of admission	Age on admission	Urea	Sodium	Potassium	Albumin	Haemoglc	White cell count	Creatinine	Urea/Creatinine	Obs died	y_hat	p_hat	Intercept	Difference
1	1	43	16.9	178	4.7	31.1	18.9	10	203.1	0.083	0	-5.87	0.003	-12.4424	0.00
1	1	69	13.6	177.3	5.8	50.3	10.9	16.2	3.8	3.579	0	4.54	0.989		
0	1	63	7.3	180.3	4.1	43.2	15	20.3	25.5	0.286	0	-4.27	0.014		
0	1	37	11.1	104	5	52.8	21.4	6.4	19.5	0.569	0	-6.44	0.002		
0	1	38	2.7	186.6	6.8	51.8	17.3	0.5	175.5	0.015	0	-8.37	0.000		
1	1	84	10.4	110.2	1.3	48.8	15	9.7	297.6	0.035	0	-4.15	0.016		
0	1	37	7.3	128.2	2.8	43.8	7.4	15.8	204.3	0.036	0	-6.02	0.002		
0	1	76	1.2	123.8	7.8	24.2	11.6	12.5	172.6	0.007	0	-4.13	0.016		
1	1	88	22.2	149.4	5.4	31.1	7.7	7.8	158.1	0.14	0	-3.06	0.045		
1	1	39	10.8	122.8	8.2	35.9	17.4	20.2	177	0.061	0	-6.07	0.002		
0	1	39	22.3	137.8	4.1	28.7	6.9	11.7	29.7	0.751	0	-3.54	0.028		
0	1	51	18.2	138.5	4	50.4	8.4	12.2	78.3	0.232	0	-5.39	0.005		
1	1	85	15.7	139.2	5.3	52.5	13.9	20.1	63.2	0.248	1	-3.52	0.029		
0	1	48	5.6	108.1	4.7	29.2	7.5	7.1	101.9	0.055	0	-5.59	0.004		
0	1	69	12.8	103.4	4.3	43.4	17.6	6.1	6.1	2.098	0	0.10	0.525		
1	1	33	13.1	146.6	2.9	37.9	12.5	8.3	112.3	0.117	0	-6.45	0.002		
1	1	43	13.6	135.2	4.6	40.6	10.9	5.9	72.1	0.189	0	-6.14	0.002		
0	1	62	11.1	168.6	8.2	31.4	21	14.6	54.6	0.203	0	-4.90	0.007		
1	1	62	12.5	103.9	3.2	46.3	10.2	21	24.6	0.508	0	-3.39	0.033		
0	1	89	0.9	174.6	1.1	22.7	6.7	4	13.1	0.069	0	-3.13	0.042		
1	1	71	22	104.5	3.2	37.7	7.6	11.7	40.7	0.541	1	-2.76	0.060		
1	1	80	4.4	138.4	2.3	32.5	7.1	4.3	116.6	0.038	0	-4.08	0.017		
0	1	56	5.7	103.4	1.4	45.6	8	9.2	41.3	0.138	0	-5.31	0.005		
1	1	56	5	163.1	6.9	17.9	11.4	12.1	116	0.043	0	-4.76	0.008		
1	1	31	21.4	187.5	7.5	19.5	7.8	2.4	47.1	0.454	0	-5.41	0.004		
0	1	47	1.2	169.1	3.3	32.3	19.5	5	53.5	0.022	1	-6.51	0.001		
0	1	70	13.9	180.4	1.1	36.7	7.3	14.9	117.5	0.118	0	-3.62	0.026		
0	1	53	7.1	162.3	4.1	48.6	16.5	13.5	185.6	0.038	0	-6.07	0.002		
1	1	25	18.3	115.7	4.6	49	12	7	94.3	0.194	0	-7.32	0.001		
1	1	78	0.1	182	6.2	35.6	21.4	15.7	161	0.001	0	-4.67	0.009		
0	1	98	6.1	178.6	7.3	45.6	21.5	13.4	149.6	0.041	0	-4.12	0.016		
0	1	97	13.2	141.5	7.8	45.8	15.1	16.4	134.5	0.098	0	-3.49	0.030		
1	1	83	5.9	107.2	8	39.5	8.4	17.9	107.8	0.055	0	-3.86	0.021		
1	1	78	7.3	152.7	2	43	15.7	6.5	106.1	0.069	0	-4.70	0.009		
1	1	24	5.3	161.8	7.7	24.3	17.9	5.9	175.2	0.03	0	-7.47	0.001		
0	1	22	7.7	183.3	6.9	29.2	10.5	23.5	63.2	0.122	0	-6.09	0.002		
1	1	28	13.5	106.8	6.5	31.7	10.1	9.6	196	0.069	0	-6.61	0.001		
0	1	60	9.6	139.8	4.8	32.8	5.3	3.9	119.5	0.08	0	-5.14	0.006		
1	1	74	3.9	140.1	5.2	14.9	18.6	7.4	207.6	0.019	0	-4.05	0.017		
1	1	91	18.2	135.7	3.8	48.1	9.1	9.8	180.9	0.101	0	-3.60	0.027		
0	1	60	12	125.6	5	25.5	5.5	23.8	47.2	0.254	0	-3.03	0.046		
0	1	100	16.9	177	6.5	41.9	16	0.5	170.6	0.099	0	-4.05	0.017		
1	1	44	0.9	159.4	1.4	26.6	12	0.2	99.3	0.009	0	-6.25	0.002		
1	1	83	9.8	120.6	5.7	46.6	5.9	19.3	49.3	0.199	0	-3.38	0.033		
1	1	82	2.6	166.5	6.5	43.4	8.5	16.3	14.9	0.174	0	-3.94	0.019		
0	1	79	20.9	103.4	6.5	15.4	17	6.6	68.7	0.304	0	-2.93	0.050		
0	1	59	0.4	159.4	3.3	16.4	13	15.3	238.1	0.002	0	-4.07	0.017		
1	1	64	0	118.9	8.5	33.1	14.3	6.6	240.7	0	0	-5.73	0.003		
0	1	77	7.1	102.3	6.9	35.1	11.1	0.7	195.7	0.036	0	-5.04	0.006		
1	1	38	10.1	179	6.1	19.7	7.2	6	183.7	0.055	0	-5.80	0.003		

Instructions using What-If Analysis

1. Go to Data tab
2. Click on What-If Analysis
3. Choose the Goal seek option
4. Enter the values as follows
   - i. Set cell **P2**
   - ii. To value **0**
   - iii. By Changing Cell **O2**
5. Press **OK**
6. Get optimum value in **O2** cell