Reviewer's report

Title: Predicting the cumulative risk of hospitalization - model development and validation

Version: 1

Date: 10 December 2013

Reviewer: Michael Campbell

Reviewer's report:

Major
1) The title is not clear. The paper is not about the risk of being hospitalised, but the risk of death during hospitalisation.
2) It is not clear whether the authors are talking about deaths post discharge (p4 line -2) or post admission (p5 line 5)
3) They don’t explain why they are interested in post discharge deaths, or why they chose 7 days after discharge. They don’t consider that deaths in hospital may have different causes to deaths post discharge.
4) They don’t define accuracy of the model. If one assumes it is the proportion of patients correctly predicted as alive or dead, they should specify the cut-off point of the prediction probability.
5) They don’t discuss the very marked circadian rhythm they show in death rate. Why should patients admitted at 0600 and 0700 have lowest mortality and patients admitted at 1700 the highest? Is it because day cases are admitted early in the morning and very few day cases die? I would have expected cases admitted in around midnight to have higher death rates than those at 17.00, when medical cover is lowest. Have other authors found a similar timing?
6) Many authors (e.g., Campbell et al. BMJ, 2012 BMJ doi=10.1136/bmj.e1001) have shown that method of admission (elective or emergency) is a major predictor but the authors don’t consider this.
7) The modelling of age is unclear. From Table 2 it would appear to only require 1 degree of freedom. This would assume a linear relation between age and death rate, which is unlikely.

Minor Essential
1) References not consistent. Eg sometimes initials and sometime first names spelled out, sometimes first names precede second and sometimes other way around. Reference 6 incomplete
2) They omitted some well known papers on predicting hospital mortality eg Campbell (2012)

3) Table 1:
i) Why choose age ranges given? Most authors separate 0-1, and 1-4
ii) Not sensible to give s.d. of LOS. Should give range and interquartile range.
iii) First label in Table wrong. Unclear what is denominator for percentages.
v) To me it looks as if the test data are older than the training data.

4) P5 last paragraph The authors have mis-understood the Nyquist frequency. This is half the sampling frequency and represents the highest frequency that a signal can represent. Thus the rate of sampling determines the Nyquist frequency and so you can’t say, as the authors do ‘the rate of sampling must exceed the Nyquist frequency’

5) P6 I do not know how ‘logistic model reliability’ can be determined using 95% confidence interval and indeed the confidence intervals are not given anywhere.

6) P4 line -4 and P7 line 13 What is meant by ‘7 days of separation?’ – an odd phrase for ‘since discharge’ if that is what is meant.

7) Typos p5 line 2 it’s => its line 8 model=> models

Level of interest: An article of limited interest

Quality of written English: Not suitable for publication unless extensively edited

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests