Identifying admitted patients at risk of dying: a prospective observational validation of four biochemical scoring systems

Mikkel Brabrand, Torben Knudsen, Jesper Hallas

ABSTRACT

Objectives: Risk assessment is an important part of emergency patient care. Risk assessment tools based on biochemical data have the advantage that calculation can be automated and results can be easily provided. However, to be used clinically, existing tools have to be validated by independent researchers. This study involved an independent external validation of four risk stratification systems predicting death that rely primarily on biochemical variables.

Design: Prospective observational study.

Setting: The medical admission unit at a regional teaching hospital in Denmark.

Participants: Of 5894 adult (age 15 or above) acutely admitted medical patients, 205 (3.5%) died during admission and 46 died (0.8%) within one calendar day.

Interventions: None.

Main outcome measures: The main outcome measure was the ability to identify patients at an increased risk of dying (discriminatory power) as area under the receiver-operating characteristic curve (AUROC) and the accuracy of the predicted probability (calibration) using the Hosmer-Lemeshow goodness-of-fit test. The endpoint was all-cause mortality, defined in accordance with the original manuscripts.

Results: Using the original coefficients, all four systems were excellent at identifying patients at increased risk (discriminatory power, AUROC ≥0.80). The accuracy was poor (we could assess calibration for two systems, which failed). After recalibration of the coefficients, two systems had improved discriminatory power and two remained unchanged. Calibration failed for one system in the validation cohort.

Conclusions: Four biochemical risk stratification systems can risk-stratify the acutely admitted medical patients for mortality with excellent discriminatory power. We could improve the models for use in our setting by recalculating the risk coefficient for the chosen variables.

INTRODUCTION

An important part of the routine work of frontline personnel in emergency departments and admission units is to assess the risk of individual patients. However, many physicians feel inadequately trained, and prognostication is not a mandatory part of medical education. As a consequence, automated risk stratification could assist physicians attending to emergency patients. However, in a recent review, none of the risk stratification tools for use in the emergency departments and admission units attained the highest level of evidence. Several systems have been developed, but only a few have been externally validated, even though this is an important part of the development process.

Some of the existing risk stratification systems are based solely on vital signs and others on biochemical analyses. Systems
Risk stratification using biochemical data

based on vital signs require manual collection of data, whereas systems based on biochemical analyses can be automated. Data can easily be extracted from the hospital computer systems and risk stratification can be performed in an automated process.

We performed the present study with the objective of validating existing risk stratification systems that predict mortality for medical patients based solely on biochemical data. Four systems based on multiple (more than two) routinely available variables (in our setting) and not restricted to selected groups of medical patients were included.

METHODS
We performed an external validation of existing biochemical risk stratification systems by applying the coefficients and ORs reported in the original papers. Furthermore, we validated the choice of variables in the original papers by recalculating the coefficients to fit our current patient population.

Setting
Sydvestjysk Sygehus is a 460-bed regional teaching hospital in the western part of Denmark with a contingency population of 220 000. All subspecialties of internal medicine are represented.

Patients can be admitted to the medical admission unit (MAU) by their general practitioner, out-of-hours emergency medical service, outpatient clinics, emergency department and ambulance services. Two attending physicians, one in internal medicine and one in cardiology, one senior resident and two interns staff the MAU.

Design and data
We conducted a prospective observational cohort study of all patients admitted through the MAU at our hospital. All consecutive adult patients (ages ≥15 years) admitted from 2 October 2008 until 19 February 2009 (first cohort) and from 23 February 2010 until 26 May 2010 (second cohort) were included in the study.

Upon admission, a nurse recorded the vital signs and registered these along with demographic information and the primary complaint on a form. After inclusion of all patients, we extracted blood test results from the hospital computer systems. No extra biochemical analyses were added as part of this study, and only analyses ordered by the admitting doctor were included. Most patients had the following biochemical standard panel taken: haemoglobin, leukocytes, platelets, C reactive protein, sodium, potassium, creatine, urea, total calcium, glucose and albumin. Almost all patients admitted to the cardiology section had troponin, amylase and calcium, glucose and albumin. Almost all patients admitted to the cardiology section had troponin, amylase and

Statistics
The sample size was dictated by another part of the study. In brief, the sample size was calibrated to develop and validate a risk-stratification system to predict 7-day all-cause mortality.

We calculated the predicted mortality using the coefficients presented in the original papers. To assess the ability of each system to identify patients at highest risk of dying (ie, the discriminatory power), we calculated the area under the receiver-operating characteristic curve (AUROC). AUROC is a summary measure of

2

sensitivity and specificity at each possible cut-off and basically represents the probability that a patient who eventually dies will have a higher score than a patient who survives. An AUROC above 0.8 is said to represent excellent discriminatory power. The calibration was assessed using the Hosmer-Lemeshow goodness-of-fit test. The calibration assesses if the observed mortality rate matches the expected rate, derived from the scoring systems. For this test, we divided the population into deciles by expected event rate. A p value above 0.05 indicates acceptable calibration. A scoring system might show excellent discriminatory power and yet have poor calibration if, for example, it was developed on a population with low overall mortality and then applied to a population with high overall mortality.

As the predictive power would be expected to vary across populations, we calculated the AUROC of each of the original scores for patients presenting with the previously specified presenting complaints.

Finally, we attempted to optimise the models to our setting by recalculating the scoring coefficients; that is, we performed the multivariable analyses anew by using the variables included in the original models. We used the first cohort (collected from 2008 to 2009) for the development and the second cohort (collected in 2010) for validation of the recalculated coefficients.

As the Asadollahi score is a set score (ranging from 0 to 20) and not a regression formula, we initially performed a new logistic regression using our development cohort. From the coefficients derived, we assigned a score (from 1

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Demographics of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total, n=5894</td>
</tr>
<tr>
<td>Female</td>
<td>2950 (50.1%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>65 (49-77)</td>
</tr>
<tr>
<td>Length of stay (days)</td>
<td>2 (1-6)</td>
</tr>
<tr>
<td>In-hospital mortality</td>
<td>205 (3.5%)</td>
</tr>
<tr>
<td>Imminent death</td>
<td>46 (0.8%)</td>
</tr>
<tr>
<td>Admitted due to infectious disorder</td>
<td>178 (3.0%)</td>
</tr>
<tr>
<td>Admitted due to malignant disorder</td>
<td>128 (2.2%)</td>
</tr>
<tr>
<td>Admitted due to endocrine disorder</td>
<td>307 (5.2%)</td>
</tr>
<tr>
<td>Admitted due to circulatory disorder</td>
<td>1375 (23.4%)</td>
</tr>
<tr>
<td>Admitted due to pulmonary disorder</td>
<td>972 (16.5%)</td>
</tr>
<tr>
<td>Admitted due to symptoms</td>
<td>1194 (20.3%)</td>
</tr>
<tr>
<td>Admitted due to observation</td>
<td>1012 (17.2%)</td>
</tr>
<tr>
<td>Admitted due to other reasons</td>
<td>718 (12.2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Variables included in the scores and the level of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Percentage of missing</td>
</tr>
<tr>
<td>Lactate dehydrogenase</td>
<td>76.6</td>
</tr>
<tr>
<td>Bilirubin</td>
<td>75.1</td>
</tr>
<tr>
<td>Alkaline phosphatase</td>
<td>75.0</td>
</tr>
<tr>
<td>Bicarbonate</td>
<td>71.6</td>
</tr>
<tr>
<td>Alanine aminotransferase</td>
<td>68.3</td>
</tr>
<tr>
<td>Neutrophil count</td>
<td>42.1</td>
</tr>
<tr>
<td>Urea/creatinine</td>
<td>13.0</td>
</tr>
<tr>
<td>Urea</td>
<td>12.7</td>
</tr>
<tr>
<td>Albumin</td>
<td>7.5</td>
</tr>
<tr>
<td>Platelets</td>
<td>7.1</td>
</tr>
<tr>
<td>Glucose</td>
<td>6.9</td>
</tr>
<tr>
<td>White cell count</td>
<td>6.0</td>
</tr>
<tr>
<td>Creatine</td>
<td>5.8</td>
</tr>
<tr>
<td>Potassium</td>
<td>5.5</td>
</tr>
<tr>
<td>Sodium</td>
<td>5.2</td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>5.1</td>
</tr>
<tr>
<td>Haematocrit</td>
<td>5.1</td>
</tr>
<tr>
<td>Age</td>
<td>0.0</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0</td>
</tr>
<tr>
<td>Mode of admission</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Required in the score.*
to 6) to each variable and recalculated the score for both cohorts. We tested calibration according to Seymour et al,\textsuperscript{14} that is, we predicted the probabilities of the individual scores using logistic regression analysis and calculated the Hosmer-Lemeshow goodness-of-fit test.

Data are reported as median (IQR) or proportions whenever appropriate. Differences between patients with and without missing data were tested using the $\chi^2$ test or Wilcoxon rank-sum test.

STATA V.12.1 (StataCorp, College Station, Texas, USA) was used for the analyses.

RESULTS

A total of 5894 patients were included in our study (see table 1 for details). Among these, 205 (3.5%) died during the admission, and 46 (0.8%) died within one calendar day.

Validation of the original scores

We could include 4925 patients (83.6% of the entire cohort) in the Prytherch score (table 2). Using the original formula, we found an AUROC of 0.842 (95% CI 0.818 to 0.865; figure 1 and table 3) and goodness-of-fit test, $\chi^2$=419.63 (10 degrees of freedom), $p<0.001$. Thus, the Prytherch score showed a good ability to identify patients at high risk of dying, but failed in calibration, as fewer patients died than expected.

In calculating the Froom score,\textsuperscript{7} we could include only 919 patients (15.6%; table 2). Using the ORs specified in the original article, we found an AUROC of 0.862 (95% CI 0.813 to 0.910; figure 1 and table 3). As the original paper did not provide the coefficient for the intercept, we were unable to reliably assess calibration. In an attempt to reduce selection bias, Froom and Shimoni\textsuperscript{7} used imputation of the mean (by assigning the

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Performance of the model using the original coefficients and after recalculation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Score</strong></td>
<td><strong>Discriminatory power (AUROC)</strong></td>
</tr>
<tr>
<td></td>
<td>Original model</td>
</tr>
<tr>
<td>Prytherch score\textsuperscript{6}</td>
<td>0.842 (0.818–0.865)</td>
</tr>
<tr>
<td>Froom score\textsuperscript{7}</td>
<td>0.862 (0.813–0.910)</td>
</tr>
<tr>
<td>Loekito score\textsuperscript{8}</td>
<td>0.922 (0.879–0.965)</td>
</tr>
<tr>
<td>Asadollahi score\textsuperscript{11}</td>
<td>0.803 (0.776–0.829)</td>
</tr>
</tbody>
</table>

Area under receiver-operating curve (AUROC) above 0.8 represents good discriminatory power, and $p$ value for calibration above 0.05 represents good calibration.
value of 2.5 to all missing variables reduced into quartiles). Adapting this approach led to the inclusion of all 5894 patients with an AUROC of 0.814 (CI 0.788 to 0.841). Again, because of a missing coefficient for the intercept, we could not assess calibration. Thus, the Froom score was good at identifying patients at high risk, but we could not assess the level of precision.

As for the Loekito score, we could include 540 patients (9.2%; table 2). Using the reported coefficients, we found an excellent discriminatory power (AUROC=0.922, CI 0.879 to 0.965, figure 1 and table 3). Calibration failed with a goodness-of-fit test, $\chi^2=30.7$, $p=0.0007$. Thus, the Loekito score showed excellent discriminatory power but failed calibration.

We could include 4863 (82.5%) in the Asadollahi score (table 2). We found a good calibration (AUROC=0.803; CI 0.776 to 0.829; figure 1 and table 3), but could not assess it because of the construction of the score in the original article.

The predictive ability of each score varied widely with each presenting complaint; however, within each complaint, the scores more or less had identical AUROCs (table 4). Overall, malignant, endocrine and pulmonary disorders had the lowest AUROC, while infectious disorders had the highest (table 4). Some of these calculations are based on limited numbers (as indicated by the CIs).

### Recalculated coefficients

Performing the recalculation of the Prytherch score, we achieved excellent AUROCs in both cohorts as well as acceptable calibration (figure 2 and table 3). Sex, urea, sodium, haemoglobin, creatine and potassium were not significantly associated with in-hospital mortality in our material, but because they were included in the original, we kept them in the analysis.

Recalculating the Froom score, we achieved excellent AUROCs in both cohorts, but calibration failed in the validation cohort (figure 2 and table 3). Age, alkaline phosphatase, alanine aminotransferase, urea, white cell count and glucose were not significantly associated with in-hospital mortality, but were kept in the model.

When recalculating the Loekito score, we found that urea, creatine, albumin, haemoglobin and white cell count were not significantly associated with the endpoint of 1-day mortality. We achieved excellent AUROCs in both cohorts as well as almost perfect calibration (figure 2 and table 3).

When recalculating the Asadollahi score, we assigned a score of one each to haemoglobin, platelets and glucose (none of which were significantly associated with the endpoint), three to sodium, four each to age and white cell count and six to urea. AUROC was excellent in both cohorts and calibration acceptable (figure 2 and table 3).

In all four methods, the discriminatory power remained constant or improved when we compared it with the calculation based on the original coefficients and ORs.
**Selection bias**

For the Prytherch, Froom and Asadollahi scores, patients who were excluded because of missing values had the same mortality as those who were included (table 5). For the Loekito score, patients with missing data had significantly lower 1-day mortality (table 5).

**DISCUSSION**

Using four existing biochemical-based risk stratification systems, we could risk-stratify acutely admitted medical patients with excellent discriminatory power. We could only evaluate the calibration for two scores, the Prytherch score\(^6\) and the Loekito score\(^8\), which both failed. When recalculating all four scores, both discriminatory power and calibration improved, except for the Froom score,\(^7\) where calibration failed.

In the present article, we focused only on biochemical-based risk stratification systems. While systems based on vital signs can be calculated shortly after arrival, biochemical-based systems require the blood tests to be analysed first. On the other hand, for systems based only on biochemical data, interobserver or intraobserver variation is virtually eliminated. We have identified four systems with broad inclusion criteria that could potentially be used in emergency departments and MAUs. The systems included were developed in different settings, ranging from floor beds\(^6,8,11\) to a medical emergency room.\(^5\) One was internally validated using a split sample technique,\(^6\) while the others were validated in external cohorts.\(^7,8,11\) However, even if the systems were developed in a setting similar to ours and validated by the original authors, they still need to be externally validated in independent cohorts, as we now have performed, before they should be used in the clinical routine.\(^4\)

Although all four systems had acceptable discriminatory power, two systems failed in calibration. One way of correcting poor calibration is to perform a recalculation. We have carried out so by performing a multivariable logistic regression in one cohort and then validating it in another. This approach generally improved the discriminatory power and made calibration acceptable. In fact, calibration became acceptable in both systems that previously failed. After recalculation, however, calibration failed in the Froom score,\(^7\) a system for which we could not test calibration using the original formula. Our best explanation for this is differences in mortality because the Froom score\(^7\) was developed and validated in cohorts with a higher mortality than ours (5.6% vs 3.5%).

The Prytherch score\(^6\) seems to fit our setting best. The discriminatory power was excellent both before and after recalculation. Calibration failed before recalculation, but was acceptable afterwards. Most important, using our standard biochemical profile, we could include the majority of our patients. Both the Froom\(^7\) and Loekito scores\(^8\) performed better, but only marginally, and the Froom score\(^7\) failed on calibration after recalculation; we could include only a few of our patients in both scores. However, the choice of score depends on several additional factors. Some hospitals might not routinely measure all investigations required by each score (eg, albumin) and some investigations are error prone (eg, haemolysis in potassium measurements). The Asadollahi score only relies on seven
parameters and could thus be easily obtained and perhaps less expensive to report on most patients. Also, it is not significantly inferior to the other scores and might therefore be more suitable for other settings.

Our study has limitations. First, we have a substantial amount of missing data. This absence is not a major problem when calculating the Prytherch or Asadollahi score, but it was for the Froom and Loekito scores. There is no doubt that this has introduced selection bias into our study. Although that we have not been able to demonstrate any selection bias for the Prytherch, Froom and Asadollahi scores looking at our primary endpoint of mortality, we showed that patients with missing data in the Loekito score had a significantly lower mortality. An apparent explanation is that bicarbonate is part of the formula. At our institution, bicarbonate is mostly analysed as part of arterial blood gas analyses and thus primarily measured in the most critically ill patients. Patients with missing data also had a significantly shorter length of stay, but were not uniformly older or younger than patients that could be included in each score (table 5). These indications of selection biases prompt us to question the external validity and generalisability of our findings, and we see this as an indication that further studies, where the risk of selection bias is minimised, are required. Second, the Loekito score requires haematocrit (we estimated this using the haemoglobin level) and total CO2 (which we estimated using bicarbonate). However, when performing our own logistic regression analyses of both systems, we had acceptable results, proving this to be of no concern. Third, this study still represents a single centre application of the scoring systems, and the results should be evaluated with this in mind. Fourth, we run a risk of overfitting when performing recalculation. With only 26 imminent fatalities in the development cohort, overfitting is a potential risk for the Loekito score. However, our validation proves that it was not an issue. As for the other three systems, we have enough fatalities for a valid recalculation.

We have found that four risk stratification systems based on biochemical data can identify patients at an increased risk of dying, although with limited precision. The models could be improved by recalculation, but the question remains if the use of these systems will improve clinical practice. In an ideal study, patients should be randomised to either be risk-stratified by a predefined system or be managed by clinical assessment alone, and the potential improvement in treatment should be measured. This approach is a complicated setup not previously performed for any of the present systems, but is the only way to show if the implementation of the system matters.

### Contributors
MB conceived and designed the study, collected, analysed and interpreted the data and wrote the report. TK and JH conceived and designed the study and assisted with analysis and interpretation of the data and writing of the report. All authors have had full access to all data and take...
responsibility for the integrity of the data and the accuracy of the analyses. MB is the guarantor. All authors have read and approved the final manuscript.

Funding The study was funded by Sydvestjysk Sygehus, Karola Jørgensens Forskningsfond, Edith og Vagn Hedegaard Jensens Fond, AB Fonden and Johs M Klein og Hustrus Mindelegat. None of the funders have had influence on the design and conduct of the study; collection, management, analysis and interpretation of the data; or preparation, review or approval of the manuscript, as the researchers are independent from all sponsors.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

Identifying admitted patients at risk of dying: a prospective observational validation of four biochemical scoring systems
Mikkel Brabrand, Torben Knudsen and Jesper Hallas

BMJ Open 2013 3:
doi: 10.1136/bmjopen-2013-002890

Updated information and services can be found at:
http://bmjopen.bmj.com/content/3/6/e002890

These include:

References
This article cites 14 articles, 2 of which you can access for free at:
http://bmjopen.bmj.com/content/3/6/e002890#BIBL

Open Access
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See: http://creativecommons.org/licenses/by-nc/3.0/ and http://creativecommons.org/licenses/by-nc/3.0/legalcode

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

Emergency medicine (271)
Epidemiology (1987)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/