A logistic regression model to predict the probability of admission

Gary Hutson, Mike Christopher, Ben Pope and Helen Johnson

ABSTRACT

Aim To create a logistic regression equation which can be used to estimate the probability of admission, based on the work previous carried out by Cameron et al. (2014). This aims to use the score to directly predict the probability of admission from the trusts (Nottinghamshire University Hospitals Trust) core patient administration system.

Methods Application of the score on the main core PAS system and testing of the predictive accuracy carried out by mixed effects logistic regression and validation of the score using a range of techniques – in the main confusion matrices and ROC curves. Improvement of the model was undertaken by looking at the significance of the predictor variables (those used to inform whether someone is admitted or not) and multiple models compared to arrive at the optimum model to use.

Results Four models were built and compared. Model 1 was based on the composite predictor variables and attained an accuracy of predicting admission of 71.3% (other measures of predictive accuracy are contained in the paper), based on comparison of other predictor variable fits, the current best performing model is entitled Model 4 – which has the inclusion of three additional variables, alongside the total score, these are ED LOS, patient admitted again within 28 days and where the patient's sex was equal (=) to male. This improved the accuracy by 1.6%.

Conclusions Change of paradigm from the focus on the score only to predicting the probability of admission; thought given to additional variables needed and iterative additional testing to try and improve the accuracy further.

INTRODUCTION

Due to the rise in Emergency Department attendances locally and nationally and the effect they have on longer waiting times and over-crowding the need and focus must be on novel and new approaches to detect and allocate resources to the patients with the highest probability and likelihood of admission into department. The paper written by Cameron et al. (2014) suggests that determining these patients at the point of triage would allow for more effective triage and decision making operationally.

This paper is a consequence of their analysis into the application of a simple score to predict admission focussing on age, EWS (Early Warning Score), triage category, referred by GP, arrived by ambulance and admitted in the previous year. This has then been simplified by us to create a Total Admission Score (TAS). This paper aims to answer the questions posed in the following Methods Section.

METHODS

Aim and Design

Application of the score to Nottinghamshire University Hospitals' core Patient Administration System (PAS) to determine whether someone will be admitted or not, as this is a binary/dichotomous variable, the need for a logistic regression model to be utilised to arrive at a regression equation to predict probabilities, or to classify into patients who are to be admitted and those who should be discharged. Benchmarking and comparison of model(s) utilised to try and improve the predictive power and accuracy of the models.

The main aim was to answer the key questions posed by the lead clinicians; these were to "***increase the accuracy of the score to be more focused to predicting admission"; "increasing the accuracy of the score at predicting admission"*** and to "***improve the score to be 95% accurate at predicting admissions***".

Variables

Predicted variable Each attendance is assigned with a flag of admission or discharge, on the PAS system, and this was then used as the predicted (dependent) variable, alongside the predictor (independent) variables.

Predictor variables A number of key variables were considered, with help of the clinical decision making team, these were built into a SQL query for multiple models to

be determined and compared. The model with the highest predictive accuracy, hitherto, contains the Total Admission Score (TAS) derived by the research from Cameron et. Al (2014) research; length of time in the ED department, returned within 28 days and patient sex = male.

Treatment of missing data and sources of bias

The data was extracted from the core PAS system and inevitably there would be some data entry issues. All duplicates were removed.

The main data omissions were due to the absence of observations, which are utilised to make up the score. Out of the attendances (n=94,488) analysed around 29% had a missing total score, due to various observations that had not been recorded on the system. Meaning the attendances (n=70,455) remained to be analysed. In addition, the data range only stems back to 12th August 2016, so this is the data used to derive the test and training samples for introduction into the model.

Statistical Analysis

All statistical analysis was carried out in the R statistical programming language, V.3.2.5. The attendances were automatically pulled from the live PAS system and assigned into two groups, one for training the model and another for testing the model fit.

The process of assignment selected was to use stratified random sampling of the patients and the apportioned sample sizes were selected as 70% for the training sample and 30% for the test sample. 10 fold cross validation was also used to try and improve the accuracy of the model and to see how it will perform with future information which is fed into the model. Only variables with a p value of <0.05 were utilised for the analysis.

RESULTS

Dataset

In total 70,455 attendances were used in the modelling of the data. The only exclusions, due to missing data, were due to missing observations (n=24,033). This missing data equates to a reduction in the size of the original sample of 29%. Data was pulled for all attendances between 12/08/2016 and 16/08/2017, giving just over a year of retrospective records to feed into the model.

Model comparisons

To improve the accuracy of the model, when applied to the live system, four separate models were compared to see if the difference in predictive power and accuracy. The focus, hitherto, has been to improve the score, but what the aim of this paper is to change the paradigm and utilise the logistic regression model, as intended, to produce a probability of admission/discharge, and have a simple and easy to understand strata such as 'high probability of admission' vs 'low probability of admission'. Cameron et al. (2014) advocated this approach *"However, the score is unlikely to be at its most useful as a simple binary predictor. Defining high probability or low probability groups might be more clinically helpful."*

In spite of either approach, it has been acknowledged that the accuracy achieved on the system currently is not sufficient enough to be utilised in practice (accuracy of 69%, or 0.694) in predicting admission, meaning that there is still a 31% chance that even if we assign a high probability of admission, you still have this probability of being discharged.

MODEL 1 RESULTS

Testing of the total score would not have been as informative as its composite parts, so each variable which makes up the Total Score was used in the regression. The relative statistics can be seen in Table 1:

Table 1 Results of univariate logistic regression for Model 1

Predictor Variable	Coefficient	OR	St Error	z value	Pr	Lower 95%	Upper 95%	Sig
(Intercept)	-3.18403	0.04142	0.03858	-82.533	0.0000	0.03839	0.04466	***
Age Score	0.25666	1.29261	0.00484	53.045	0.0000	1.28042	1.30494	***
Triage Score	0.13205	1.14116	0.00341	38.725	0.0000	1.13359	1.14884	***
Referral Score	0.02714	1.02751	0.00474	5.720	0.0000	1.01800	1.03711	***
Admitted in previous 12 months	0.08394	1.08757	0.00465	18.058	0.0000	1.07705	1.09752	***
AVPU Score	0.11073	1.11709	0.02284	4.847	0.0000	1.06842	1.16853	***
BP Score	0.19665	1.21732	0.01752	11.227	0.0000	1.17634	1.25996	***
Arrival Score	0.10834	1.11442	0.00470	23.071	0.0000	1.10422	1.12473	***
HR Score	0.34563	1.41288	0.01523	22.694	0.0000	1.37140	1.45577	***
Resp Score	0.26777	1.30705	0.01487	18.005	0.0000	1.26962	1.34585	***
Sats Score	0.12299	1.14339	0.01623	8.255	0.0000	1.10767	1.18044	***
Temperature Score	0.02739	1.02777	0.01260	2.174	0.0297	1.00269	1.05347	*

Significance codes 0.05 '.' 0.01 '*' 0.001 '**' 0 '***'

The Glasgow Admission paper highlighted that the score derived from their analyses was effective and significant at predicting admission, this can be seen in the sig column of Table 1, the stars indicate high significance with all the predictor variables, the exception here is the temperature score, which is significant, but to a lesser degree of magnitude.

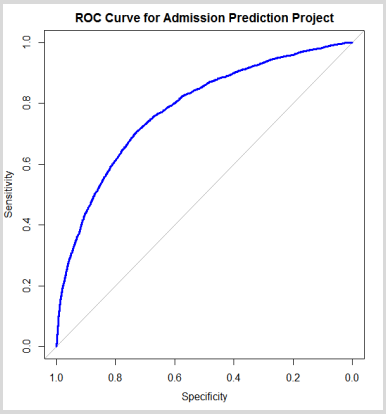
Performance of score in predicting admission

This model, when applied to the patients of this trust is accurate 71.3% of the time. The key comparison metrics are laid out in the confusion matrix, alongside a ROC curve in Table 2.

Figure 1 Comparing Model 1 results

Prediction	Actual		
	Actual Discharge	Actual Admission	Total
Predicted Discharge	7,120	3,019	10,139
Predicted Admission	3,047	7,950	10,997
Total	10,167	10,969	21,136

Accuracy	71.3%
Misclassification rate	28.7%
Sensitivity	72.5%
Specificity	70.0%
Precision	72.3%
Prevalence	51.9%
False positive rate	30.0%



The ROC curve shows that the base model derived from the admission prediction score has predictive potential (the hyperplane through the middle would be indicative of a model with no predictive power). A key statistic is 72.5% sensitivity, which means when the model predicts that a patient will be admitted they are actually admitted. The precision statistic of 72.3% shows that the when the model predicts an admission it gets it right 72.3% of the time.

The Glasgow Admission Paper shows an AUC of 87.7%, which means that their model was more accurate, conversely their approach to the research method is markedly different from this application of the model, which is used as an early warning indicator from the live patient attendance data. Their study focussed on subsets of patients who they knew were at the point of triage and their dataset was larger than the data used in this model (total records 322,846 compared to 70,455 attendances).

The only difference between Model 1 and Model 2 was the exclusion of the temperature variable, due to its lesser significance highlighted in Table 1 and the respective models Akaike information criterion (AIC) showed no change in value when this predictor variable was removed.

From the first two models it was acknowledged that new predictor variables needed to be added to the model to enhance the accuracy of the prediction. The new variables entered into Model 3 were patient returned in 7 days; patient returned in 28 days; Emergency Department LOS (Length of Stay) and the sex of the patient. Out of these additional variables returned in 28 days, ED LOS and patient sex transpired to be statistically relevant to the model. These were retained and used in the testing of Model 4. This will be the focus of the analysis below and it is the best performing model, to date, when applied to the NUH ED data.

MODEL 4 RESULTS

The end results of the final model show an improvement, but this is only minimal compared to model one. The accuracy estimates have increased by 1.6%, which is not a massive gain, but could be improved further with additional variables tested.

Table 2 Results of univariate logistic regression for Model 4

Predictor Variable	Coefficient	OR	St Error	z value	Pr	Lower 95%	Upper 95%	Sig
Intercept	-3.70700	0.02455	0.04014	-92.349	0.0000	0.02269	0.02655	***
Total Admission Score	0.12320	1.13108	0.00165	74.765	0.0000	1.12745	1.13475	***
ED LOS	0.00488	1.00490	0.00009	55.229	0.0000	1.00472	1.00507	***
Patient Returned within 28 days	-0.20980	0.81077	0.02993	-7.009	0.0000	0.76456	0.85974	***
Gender Male	0.04417	1.04516	0.02113	2.091	0.0366	1.00276	1.08934	*
Significance codes	0.05 '.' 0.01 '*' 0.001 '**' 0 '***'							

The confusion matrix shows an improvement from Model 1, by reducing the independent and/or composite scores into the total admission score and adding the ED LOS, patient returned within 28 days and testing the gender types (of which male attendees are significant, p value of 0.036).

Figure 2 Comparing Model 4 results

Prediction	Actual		
	Actual Discharge	Actual Admission	Total
Predicted Discharge	7,431	2,993	10,424
Predicted Admission	2,741	7,971	10,712
Total	10,172	10,964	21,136
Accuracy	72.9%		
Misclassification rate	27.1%		
Sensitivity	72.7%		
Specificity	73.1%		
Precision	74.4%		
Prevalence	51.9%		
False positive rate	26.9%		

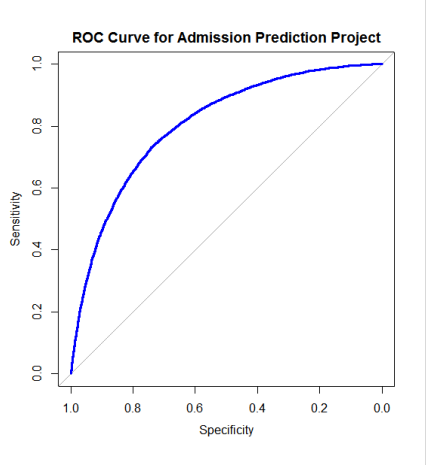


Figure 2 shows an improvement from Model 1 in most of the key prediction statistics. The accuracy increases by 1.5%, the sensitivity improves by 72.7% which relates to how sensitive the model is at predicting the correct outcome regarding admission, specificity improves by 3.1% - this relates to how accurate it is at predicting discharge when the patient is discharged. False positives and misclassifications also reduce, under the proposed model.

Cross validation (10 fold) was also used to test all four models, but did not offer any gain in performance over the stratified random sampling approach.

Using the regression equation to predict admission

The regression equation, derived from Model 4, could be utilised to predict the probability of someone being admitted and you even classify if some is likely to be admitted, or not. The probability option could have suitable cut offs such as < **(less than) 25% probability = likely to be discharged**, **50% probability equal probability of being admitted or discharged** and > **(greater than) 75% likely to be admitted**. These could be refined to add an upper banding such as >90% highly likely to be discharged. This is something to consider off the back of this analysis.

The equation to use would be plugged into the logistic regression formula to obtain the associated probabilities and would take the form of:

$$\hat{p} = \frac{\exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}{1 + \exp(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}$$

The values to substitute into the equation would be:

- Intercept coefficient for beta zero (b0) – for model four this is -3.70700.
- Total admission score coefficient (b1) – this is 0.12320.
- Actual admission score assigned to the attending patient (X1).
- ED LOS coefficient (b2) – this is 0.00488.
- Patients length of time in the department (X2) – this would be useful as the metric is recorded in minutes, so if it was built into a live system it could be refreshed every minute to refine the estimated probability of admission. It seems axiomatically obvious that if someone is in the ED department for over an hour they are more likely to be admitted, but this would be a way of quantifying this subjective judgement.
- Patient returned within 28 days coefficient (b3) – this is -0.20980 (the negative symbol indicates that this variable is more likely to estimate the odds of discharge over admission).
- Whether the patient returned within 28 days – coded as 1 for return or 0 for did not return (X3).
- Patient gender = male coefficient – this is 0.04417 (b4).
  Indicates whether the sex of the patient is male (coded as 1 for male and 0 for female) – so this is the presenting sex of the patient attending (X4).

CONCLUSION AND RECOMMENDATIONS

This paper has focused on trying to use the Glasgow Admission Score, alongside local system variables, to produce a regression equation that can then be used on a live system to estimate the probability of admission.

From the analysis conducted there has been an improvement by utilising additional variables alongside the score to improve the accuracy of the estimated probability of admission. This still does not achieve a 95% accuracy rate, so the consideration needs to be given to additional predictor variables to be used, in the model, to try and improve accuracy. Attempts to weight the score, used in Model 1, have yielded diminished model accuracy estimates, so changing the weighting of the score has a negative return on the accuracy of the prediction.

The recommendations stemming from this are to work with key clinicians in ED to understand some additional variables which can be built into the dynamic SQL query to improve the accuracy; to make sure that the additional variables can be updated in real time and can be transferrable to other trusts and that these variables are tested in the same fashion and method – therefore, utilising the same techniques used in this aforementioned paper.

Furthermore, more work is needed, if we are to boost the accuracy of this score/prediction to the required threshold. Indeed the gauntlet remains.

REFERENCES

1 Cameron A, et al. Emerg Med J 2015; 32:174–179.