

Using machine learning to predict paediatric 30-day unplanned hospital readmissions: a case-control retrospective analysis of medical records, including written discharge documentation

Huaqiong Zhou^{1,2} RN, BSc(Nurs), MCN(PaedNurs), PhD Candidate, Clinical Nurse, Research Officer

Matthew A. Albrecht² BSc(Hons), PhD, Research Fellow

Pamela A. Roberts² RN, BAppSc(Nurs), GradDipBus, MHLthAdmin, PhD, Adjunct Professor

Paul Porter^{2,3} MBBS, FRACP, Paediatrician

Philip R. Della^{2,4,5} AM, RN, RM, BAppSc(Nurs), GradDipMan, MBus, PhD, FACN, John Curtin Distinguished Professor of Nursing, Head of School

¹General Surgical Ward, Princess Margaret Hospital for Children, Perth, WA 6008, Australia.

²School of Nursing, Curtin University, GPO Box U 1987, Perth, WA 6845, Australia.

Email address: h.zhou@curtin.edu.au; matthew.albrecht@curtin.edu.au; p.a.roberts@curtin.edu.au; paul.porter@curtin.edu.au

³Joondalup Health Campus, Joondalup, WA 6027, Australia.

⁴Visiting Professor, College of Nursing, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China.

⁵Corresponding author. Email: p.della@curtin.edu.au

Abstract.

Objectives. To assess whether adding clinical information and written discharge documentation variables improves prediction of paediatric 30-day same-hospital unplanned readmission compared with predictions based on administrative information alone.

Methods. A retrospective matched case-control study audited the medical records of patients discharged from a tertiary paediatric hospital in Western Australia (WA) between January 2010 and December 2014. A random selection of 470 patients with unplanned readmissions (out of 3330) were matched to 470 patients without readmissions based on age, sex, and principal diagnosis at the index admission. Prediction utility of three groups of variables (administrative, administrative and clinical, and administrative, clinical and written discharge documentation) were assessed using standard logistic regression and machine learning.

Results. Inclusion of written discharge documentation variables significantly improved prediction of readmission compared with models that used only administrative and/or clinical variables in standard logistic regression analysis ($\chi^2_{17} = 29.4$, $P = 0.03$). Highest prediction accuracy was obtained using a gradient boosted tree model (C-statistic = 0.654), followed closely by random forest and elastic net modelling approaches. Variables highlighted as important for prediction included patients' social history (legal custody or patient was under the care of the Department for Child Protection), languages spoken other than English, completeness of nursing admission and discharge planning documentation, and timing of issuing discharge summary.

Conclusions. The variables of significant social history, low English language proficiency, incomplete discharge documentation, and delay in issuing the discharge summary add value to prediction models.

What is known about the topic? Despite written discharge documentation playing a critical role in the continuity of care for paediatric patients, limited research has examined its association with, and ability to predict, unplanned hospital readmissions. Machine learning approaches have been applied to various health conditions and demonstrated improved predictive accuracy. However, few published studies have used machine learning to predict paediatric readmissions.

What does this paper add? This paper presents the findings of the first known study in Australia to assess and report that written discharge documentation and clinical information improves unplanned rehospitalisation prediction accuracy in a paediatric cohort compared with administrative data alone. It is also the first known published study to use machine learning for the prediction of paediatric same-hospital unplanned readmission in Australia. The results show improved predictive performance of the machine learning approach compared with standard logistic regression.

What are the implications for practitioners? The identified social and written discharge documentation predictors could be translated into clinical practice through improved discharge planning and processes, to prevent paediatric 30-day all-cause same-hospital unplanned readmission. The predictors identified in this study include significant social history, low English language proficiency, incomplete discharge documentation, and delay in issuing the discharge summary.

Keywords: administrative data, clinical information, discharge planning, discharge summary, follow-up plan, machine learning, medical records, paediatric hospital readmissions, paediatric unplanned readmissions, retrospective analysis, social history, social predictors, written discharge documentation.

Received 14 April 2020, accepted 18 June 2020, published online 12 April 2021

Introduction

The identification of predictive factors associated with paediatric unplanned readmission to hospital can be used to improve discharge planning processes, and thereby help prevent such readmissions. Prior research has uncovered many of these factors; a recent systematic review¹ of the existing literature extracted 36 unique predictors associated with paediatric unplanned hospital readmissions from 44 studies. The most commonly cited four predictors were comorbidity, health insurance status, length of stay (LOS), and age at the index admission. The review highlighted that statistical identification of predictors depended on what variables were examined in each of the studies. In 33 of the 44 studies, administrative databases and medical records were both accessed. In the remaining 11 studies, only administrative variables were analysed. The number of examined variables ranged from 2² to 44³. Extracting variables from electronic or hard-copy medical records enriches the data and may assist to rectify coding errors in the administrative dataset. Manual review of medical records does, however, incur significant time and financial impost. Nevertheless, the enhanced prediction capability via including such information may result in significant reductions in readmission rate and healthcare costs.

In addition to sociodemographic and clinical information, three paediatric studies examined the association between written discharge documentation (e.g. follow-up plan or discharge summary) and unplanned readmissions, but the results were not consistent.⁴⁻⁶ In this project, written discharge documentation refers to not only the discharge summary, but also the last entry within the patient progress notes by doctors, allied healthcare providers, and nurses, as this method allows for comprehensive review of the inter-healthcare professional team members input to the discharge documentation. Variations between studies in how this information is extracted and analysed, along with its effect on prediction of readmissions, suggests further investigations are warranted. Written discharge documentation plays a critical role in the continuity of care following hospital discharge, but extracting this data is challenging for researchers.⁷⁻⁹

Apart from adding variables for predictive model development, advances in statistical analysis methods may also improve prediction accuracy, especially with large healthcare datasets. Logistic regression analysis methods are commonly employed

in predicting paediatric unplanned hospital readmissions. Advanced machine learning analysis approaches have also been applied to adult¹⁰ and paediatric¹¹⁻¹⁴ unplanned hospital readmissions, because of their potential to improve predictive model performance.¹⁵ The commonly applied approaches included random forests,¹⁶ least absolute selection and shrinkage operator (LASSO),^{11,16,17} and gradient boosted decision trees.¹⁶⁻¹⁸ However, the number of paediatric studies remains limited, and have so far only analysed administrative data.¹¹⁻¹⁴

In a recently published study,¹⁹ we developed a logistic regression model based on 16 administratively collected variables as electronic medical records were not available. The model was found with moderate discriminative ability for 30-day all-cause readmission at a tertiary paediatric hospital in Western Australia (WA) (C-statistic = 0.645).

Study aim

This current study added clinical information and written discharge documentation with the aim to determine whether adding these variables improves prediction of 30-day same-hospital unplanned readmissions compared with examining only administratively collected variables. Prediction accuracy was also examined comparing standard logistic regression analysis to machine learning approaches.

Methods

Study design

A retrospective matched case-control study was conducted, which audited the medical records of patients discharged from a tertiary paediatric facility in WA that has approximately 250 000 inpatient and outpatient visits each year.²⁰ Ethics approvals were obtained from the Human Ethics Research Committee of Health Service, Department of Health, WA (2015/55), Children's Hospital (2015015EP), and Curtin University (HR184/2015).

Data source

The patients included in this study were discharged between 1 January 2010 and 31 December 2014. The original electronic administrative inpatient dataset was extracted from the WA Hospital Morbidity Data Collection (WAHMDC). A total of 3330 patients (4.55%) experienced 30-day unplanned hospital

Table 1. Three groups of extracted variables

Administratively collected (16)	Clinical (11)	Written discharge documentation (13)
Age	Significant social history (legal custody or patient was under the care of Department for Child Protection)	Completion of <i>Nursing Admission and Discharge Planning Form</i> (Admission section and Discharge Planning section)
Sex	Language other than English	Operation sheet or the last entry progress note made by doctors
Admission status	Significant laboratory result	Clinical pathway or the last entry progress note made by nurses
Length of hospital stay (LOS)	Significant imaging result	Last entry progress note made by allied healthcare providers
Funding source as an inpatient	Significant vital signs	Written evidence of discharge information given by doctors
Health insurance status	Added new medication at discharge upon existing regular medication regime	Written evidence of discharge information given by nurses
Source of referral transport	Number of co-diagnosis recorded in the patient progress notes	Written evidence of discharge medications information by doctors
State/Territory of residence	Known allergies	Written evidence of discharge medications information by nurses
Care type	Usage of hospital services 12 months prior to the index admission: number of emergency department (ED) presentations	Written evidence of follow-up information given by doctors
Socioeconomic indexes for areas (SEIFA)	Usage of hospital services 12 months prior to the index admission: number of hospitalisations	Written evidence of follow-up information given by nurses
Distance to hospital	Usage of hospital services 12 months prior to the index admission: number of outpatient clinic attendances	Consistency of written discharge documentation among healthcare providers
Had general anaesthetic		Delay in issuing discharge summary (date of discharge summary being issued – date of discharge)
Had intensive care unit (ICU) stay		
Day of admission date		
Day of discharge date		
Number of co-diagnosis		

readmission.²¹ Hospital readmission was operationalised as an unexpected hospitalisation within 30 days as measured from an index admission. The readmission is related to the principal diagnosis of the index admission. The identification of unplanned hospital readmissions in this study was based on the combination of admission type (emergency) and the principal diagnosis of the subsequent admission following the index admission. Because of the burden associated with extracting data from medical records, out of the initial dataset, 550 patients with readmissions were randomly selected and matched to 550 patients without readmissions by age, sex, principal diagnosis of the index admission, and proportion of principal diagnosis. The randomisation and matching was generated using Coarsened Exact Matching.²² Due to the unavailability of medical records for some patients, the final number of paired patients was 470 (total patients = 940).

Sample size

Sample size was calculated based on the association between written discharge documentation and unplanned paediatric readmissions. Previous research⁵ found the absence of a written discharge plan demonstrated an odds ratio (OR) of 1.55 for readmissions. Other substantive predictive variables, such as comorbidity, possessed ORs from 1.18 to 5.61.^{2,23,24} Therefore, we consider OR for written discharge documentation to be suitable for a baseline power calculation. Assuming a rate of 40% written discharge absence/incompleteness from the larger data set, we would need 332 matched case-control pairs (with continuity correction; total = 664, for power = 0.8, and $\alpha = 0.05$) assuming the equal proportion of rehospitalisations in

each group.²⁵ Given our current sample size of 940, we have the power to detect a variable with an OR of 1.45.

Machine learning methods can sometimes require a substantially larger sample size. We therefore used multiple machine learning methods with specific reference to methods that use strong regularisation (e.g. the elastic net) – recommended for situations with a high variable to sample size ratio – and looked for consistency across algorithms for identifying important variables.

Extracted variables

Three groups of variables were analysed (Table 1). The first group (16 administrative variables) were extracted from the initial electronic dataset; the second group (11 clinical information variables) were extracted from patients' medical records; and the third group (13 variables on written discharge documentation) were extracted from the last written entry of healthcare providers in patient progress notes and/or from clinical care pathway. The data extraction was completed by HZ, using a data collection form to ensure consistency. PRD was consulted with any queries. The written discharge documentation variables were initially extracted from patients' medical records and then categorised as 'Yes/No/Not Applicable'. In particular, the *Nursing Admission and Discharge Planning Form* consists of multiple entry areas to be recorded (this form is divided in two sections, Admission and Discharge Planning); our categorisation of 'completeness' was made when all areas of the form were recorded. Partially recorded forms were considered 'incomplete'. The filled contents of the form were extracted and assessed against variables of 'Significant

social history (legal custody or patient was under the care of the Department for Child Protection)', 'language spoken other than English', 'known allergies', 'discharge information', 'discharge medication information', and 'follow-up information'.

Missing data

The numbers of missing values were as follows: Significant social history (0 without readmission, 1 with readmission); Source of referral transport (55 without readmission, 59 with readmission); and Completeness of *Nursing Admission and Discharge Planning Form* (6 without readmission, 14 with readmission). Missing data were imputed by random forest imputation using the *missForest* package in R.²⁶ This method performs well compared with other imputation procedures, and is able to impute continuous and categorical data, and allows for interactive and non-linear effects. We used default parameter settings from *missForest* (number of trees = 100, and max iterations = 10).

Statistical analysis

Data processing and analyses were conducted in R (version 3.5.1).²⁷

Model comparison of the three sets of variables

This study was interested in whether a group of variables improved prediction, and, to reduce the number of comparisons, we compared three groups of variables by sequentially fitting three logistic regression models: (1) Administrative variables only; (2) Administrative and clinical variables; (3) Administrative, clinical, and written discharge documentation variables.

Analysis of deviance with Chi-squared (χ^2) test was used for determining significance. Analysis of individual variables was not conducted at this stage, but is included in [Table 2](#) for comparison. To complement the logistic regression we used machine learning to highlight variables of relevance for prediction.

Prediction models

Multiple methods were used to ensure consistency and robustness across models, and included logistic regression, stepwise logistic regression, random forest, elastic net, and gradient boosted trees. Performance was evaluated using the C-statistic across the ten repeats of the ten-fold cross-validation.

Stepwise regression methods are standard selection methods in the relevant, existing literature. The 'glmStepAIC' method within the 'caret' package²⁸ was used for forward stepwise selection to the logistic regression model with the Akaike information criteria (AIC) penalty. Backward elimination gave the same results as forward elimination; therefore, only forward elimination is reported.

Elastic net mixes two regression penalty methods: least absolute shrinkage and selection operator (LASSO)²⁹ penalty, and ridge penalty.³⁰ It provides stable and sparse estimates of model parameters. The LASSO penalty produces sparse predictor matrices by shrinking variables, with a proportion shrunk to 0. The ridge penalty shrinks smoothly all coefficients towards 0, while retaining all variables in the model. We used the 'glmnet' package within 'caret' to perform the elastic net.

Optimal parameters were evaluated using grid search (α and λ between 0 and 1, with 0.02 step increments).

Random forests build multiple decision trees to create a 'forest' of trees. Each tree is built on a bootstrapped sample of the training data and, at each split, a random subset of the features are chosen for prediction. The number of variables randomly sampled at each split ranged from 2 to 10, in steps of 2. We used the 'randomForest'³¹ implementation within 'caret'.

Gradient boosted decision trees are similar to random forests. Trees are iteratively grown using the outcomes from a previously grown tree, applying a larger weighting to the errors from the previous tree's classifications. The 'xgboost' implementation³² within 'caret' was used. The following tuneable parameters were determined by grid search: interaction depth (from 1 to 5), fraction of variables randomly sampled for each tree (0.1, 0.2, 0.5), and minimum loss reduction to make a split $\gamma = 3, 5, 7$. The learning rate $\eta = 0.01$, and number of trees = 500.

Multiple methods were used to ensure consistency and robustness across models, and included logistic regression, stepwise logistic regression, random forest, elastic net, and gradient boosted trees. These specific methods were selected primarily because they represent the most commonly used methods in the current hospital readmission literature.

Variable selection

For models with in-built selection (stepwise regression, gradient boosted tree, and elastic net), variable selection was done through the model fitting procedure. For the random forest, we selected the top ten variables according to their variable importance. Supplementary Table S1 presents the relative variable importance for the random forest algorithm. The built-in 'varImp' function from the 'caret' package was used to calculate importance. Variable importance quantifies the relative contributions of each variable to the model, defined as the number of times a variable is selected for splitting, weighted by the improvement to the model, and averaged.

Results

Patients' characteristics, based on the three groups of variables, for the with-readmission group and without-readmission group, are presented in [Table 2](#). The length of the index admission (mean \pm s.d.) was longer in the with-readmission group compared with the without-readmission group (3.3 ± 6.6 vs 3.0 ± 6.9 days). Patients with significant social history were almost doubled in the with-readmission group compared with the without-readmission group (52 (11.1%) vs 8 (6.0%)). Five patients in the with-admission group required interpreter service but none in the without-readmission group required this service. The mean length of delay in issuing a discharge summary was longer in the with-readmission group compared with the without-readmission group (22.9 ± 39.9 vs 16.8 ± 34.3 days).

Comparison of administrative, administrative and clinical, and administrative, clinical, and written discharge documentation variable groups

The improvement in prediction of unplanned hospital readmissions for each set of variables (administrative, administrative and clinical, and administrative, clinical, and written discharge

Table 2. Characteristics of patients with readmission and without readmission
 Data are presented as mean±s.d. or n (%) unless otherwise noted. LOS, length of stay; SEIFA, Socioeconomic Indexes for Areas; ICU, intensive care unit; ED, emergency department

Variables	Patients with readmission n = 470		Matched patients without readmission n = 470		χ ² /z-test	d.f.	P-value
	Mean ± s.d. n (%)	Median (10th, 90th percentile)	Mean ± s.d. n (%)	Median (10th, 90th percentile)			
Administrative							
Age	5.2 ± 4.7	3 (0, 13)	5.3 ± 4.7	4 (0, 13)	0.36	938.0	0.72
LOS at index admission	3.3 ± 6.6	1 (0, 7)	3.0 ± 6.9	1 (0, 5.1)	0.63	936.4	0.53
SEIFA percentile	64 ± 28	68.5 (22, 96.5)	62 ± 27	67 (20, 95)	0.50	927.9	0.62
Distance from residential address to hospital	63 ± 135	18 (7, 98)	75 ± 148	18 (7, 238)	1.25	921.5	0.21
No. co-diagnoses	2.3 ± 1.9	2 (1, 4)	2.2 ± 2.0	2 (1, 4)	0.92	930.7	0.36
Sex (female/male)	272 (57.9)/198 (42.1)		272 (57.9)/198 (42.1)		0.00	1.0	1.00
Admission status (elective/emergency)	368 (78.3)/102 (21.7)		373 (79.4)/97 (20.6)		0.10	1.0	0.75
Funding source as an inpatient (Medicare/private health insurance/other)	424 (90.2)/9 (1.9)/37 (7.9)		427 (90.9)/13 (2.7)/30 (6.4)		1.50	2.0	0.48
State/Territory of residence (Western Australia/other)	468 (99.6)/2 (0.4)		466 (99.1)/4 (0.9)		0.17	1.0	0.68
Insurance status (private/public)	245 (52.1)/225 (47.9)		250 (53.2)/220 (46.8)		0.07	1.0	0.79
Source of referral transport (ambulance/aeromedical/private transport)	338 (71.9)/2 (0.4)/71 (15.1)		338 (71.9)/6 (1.3)/71 (15.1)		1.99	2.0	0.37
Care type (acute care/other)	462 (98.3)/8 (1.7)		464 (98.7)/6 (1.3)		0.07	1.0	0.79
Had general anaesthetics (no/yes)	271 (57.7)/199 (42.3)		272 (57.9)/198 (42.1)		0.00	1.0	1.00
Had ICU stay (no/yes)	460 (97.9)/10 (2.1)		465 (98.9)/5 (1.1)		1.08	1.0	0.30
Day of admission date (Monday to Sunday)	75 (16.0)/55 (11.7)/92 (19.6)/75 (16.0)/65 (13.8)/54 (11.5)/54 (11.5)		73 (15.5)/69 (14.7)/83 (17.7)/65 (13.8)/68 (14.5)/64 (13.6)/48 (10.2)		4.05	6.0	0.67
Day of admission date (weekday/weekend and public holiday)	351 (74.7)/119 (25.3)		355 (75.5)/115 (24.5)		0.05	1.0	0.82
Day of discharge date (Monday to Sunday)	59 (12.6)/67 (14.3)/69 (14.7)/86 (18.3)/79 (16.8)/52 (11.1)/58 (12.3)		77 (16.4)/72 (15.3)/78 (16.6)/65 (13.8)/76 (16.2)/49 (10.4)/53 (11.3)		6.41	6.0	0.38
Day of discharge date (weekday/weekend and public holiday)	350 (74.5)/120 (25.5)		358 (76.2)/112 (23.8)		0.28	1.0	0.60
Clinical							
No. hospital admissions in the previous 12 months	1.2 ± 1.3	1 (0, 2)	1.0 ± 0.7	1 (0, 1)	4.82	535.7	<0.01
No. ED presentations in the previous 12 months	0.8 ± 2.7	0 (0, 2)	0.1 ± 0.7	0 (0, 0)	4.31	712.5	<0.01
No. outpatient clinic attendances in the previous 12 months	2.3 ± 7.4	0 (0, 4)	0.9 ± 5.5	0 (0, 1)	3.37	861.4	<0.01
No. significant laboratory results	1.6 ± 3.5	0 (0, 7.2)	1.5 ± 3.4	0 (0, 5)	0.50	937.6	0.62
No. significant imaging results	1.0 ± 2.2	0 (0, 3)	0.9 ± 2.2	0 (0, 3)	0.92	929.4	0.36
No. significant vital signs	1.3 ± 1.9	0 (0, 4)	1.2 ± 1.8	0 (0, 3)	1.35	934.3	0.18
No. past medical history recorded in the progress notes	1.1 ± 1.7	1 (0, 3)	0.6 ± 1.0	0 (0, 2)	5.64	776.2	<0.01
Significant social history (no/yes)	417 (89.0)/52 (11.0)		442 (94.0)/28 (6.0)		7.28	1.0	<0.01
Language spoken other than English (no/yes/interpreter service required)	456 (97.0)/9 (2.0)/5 (1.0)		460 (97.9)/10 (2.1)/0 (0)		5.07	2.0	0.08
Added new medication at discharge upon regular medications (no/yes)	438 (93.1)/32 (6.9)		445 (94.7)/25 (5.3)		0.67	1.0	0.41
History of allergies (no/yes)	428 (91.1)/42 (8.9)		422 (89.8)/48 (10.2)		0.31	1.0	0.58

Written discharge documentation	22.9 ± 39.9	8 (0, 68)	16.8 ± 34.3	6 (0, 41)	2.38	844.2	0.02
Delay of issuing discharge summary (days)	232 (49.4)/224 (47.7)		271 (57.7)/193 (41.1)		4.96	1.0	0.03
Completeness of Nursing Admission and Discharge Planning Form, Admission section (yes/no)	218 (46.4)/238 (50.6)		272 (57.9)/192 (40.9)		10.40	1.0	0.01
Completeness of Nursing Admission and Discharge Planning Form, Discharge Planning section (yes/no)							
Progress note or operation sheet last entry by doctors (yes/no)	454 (96.6)/16 (3.4)		452 (96.2)/18 (3.8)		0.03	1.0	0.86
Progress note or Clinical Pathway last entry by nurses (yes/no)	426 (90.6)/44 (9.4)		441 (93.8)/29 (6.2)		2.90	1.0	0.09
Progress note last entry by allied healthcare providers (yes/no)	371 (78.9)/99 (21.1)		389 (82.8)/81 (17.2)		1.99	1.0	0.16
Discharge information documented by doctors (yes/no)	454 (96.6)/16 (3.4)		456 (97.0)/14 (3.0)		0.03	1.0	0.85
Discharge information documented by nurses (yes/no)	431 (91.7)/39 (8.3)		444 (94.5)/26 (5.5)		2.90	2.0	0.23
Discharge medication information documented by doctors (yes/no/not applicable)	277 (58.9)/188 (40.0)/5 (1.1)		262 (55.7)/202 (43.0)/6 (1.3)		1.01	2.0	0.60
Discharge medication information documented by nurses (yes/no/not applicable)	262 (55.7)/186 (39.6)/22 (4.7)		253 (53.8)/204 (43.4)/13 (2.8)		3.30	2.0	0.19
Follow-up plan documented by doctors (yes/no/not applicable)	436 (92.8)/21 (4.5)/13 (2.8)		433 (92.1)/16 (3.4)/21 (4.5)		2.57	2.0	0.28
Follow-up plan documented by nurses (yes/no/not applicable)	387 (82.3)/70 (14.9)/13 (2.8)		382 (81.3)/67 (14.3)/21 (4.5)		1.98	2.0	0.37
Consistency of written discharge information documented among healthcare providers (yes/no)	364 (77.4)/106 (22.6)		376 (80.0)/94 (20.0)		0.77	1.0	0.38

Table 3. Model performance comparison

Analysis approaches	C-statistics/ Mean	C-statistics/s.d.	Sensitivity	Specificity
Model 1: Administrative model				
Logistic regression	0.487	0.066	0.506	0.465
Stepwise logistic regression	0.477	0.049	0.531	0.444
Random forest	0.519	0.061	0.533	0.503
Elastic net	0.5	0	1	0
Gradient boosted tree	0.509	0.045	0.507	0.507
Model 2: Administrative + clinical model				
Logistic regression	0.585	0.051	0.653	0.468
Stepwise logistic regression	0.593	0.058	0.708	0.451
Random forest	0.603	0.054	0.679	0.46
Elastic net	0.616	0.047	0.836	0.358
Gradient boosted tree	0.624	0.054	0.869	0.319
Model 3: Administrative + clinical + written discharge documentation model				
Logistic regression	0.609	0.054	0.646	0.514
Stepwise logistic regression	0.617	0.05	0.654	0.512
Random forest	0.642	0.052	0.652	0.521
Elastic net	0.635	0.048	0.753	0.414
Gradient boosted tree	0.654	0.053	0.774	0.439

documentation) was sequentially assessed using logistic regression model comparison with standard significance testing. A model with only administrative variables did not significantly improve prediction (administrative model vs intercept only model, $\chi^2_{32} = 27.4, P = 0.70$). By contrast, the inclusion of clinical variables significantly improved prediction over the administrative-only model ($\chi^2_{12} = 86.1, P < 0.01$), and the inclusion of written discharge documentation variables further improved prediction over the administrative and clinical variables model ($\chi^2_{17} = 29.4, P = 0.03$).

Prediction model performance of standard logistic regression to machine learning approaches

Prediction performance for each method obtained from the 10 × 10-fold cross-validation is presented in Table 3. The best performing prediction model according to the mean receiver operating curve (ROC) statistic (C-statistic) was the gradient boosted tree model using all three sets of variables (administrative, administrative and clinical, and administrative, clinical, and written discharge documentation), followed closely by the random forest and elastic net. Consistent with the logistic regression above, models using only administrative data performed no better than chance, and substantial improvements in the C-statistic were seen by including clinical and written discharge documentation data.

Fig. 1 presents the ROC curves for each machine learning algorithm. ROC curves were extracted from the predictions of the 10 × 10-fold cross-validation.

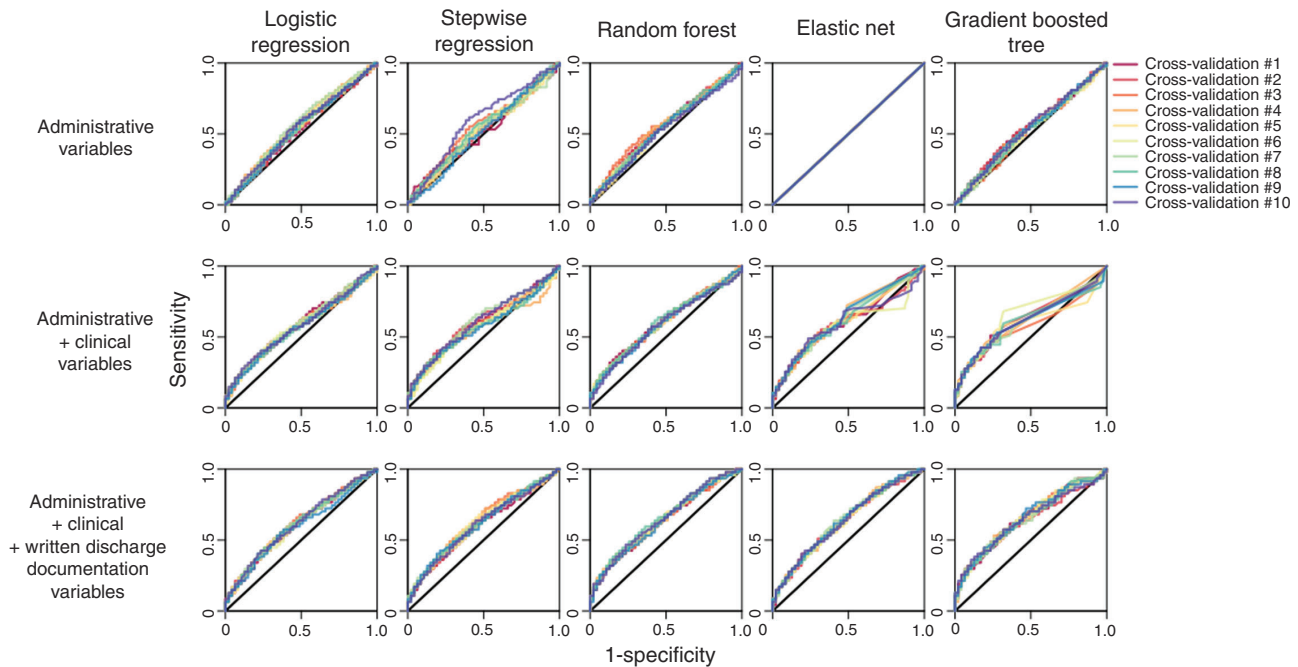


Fig. 1. Receiver operating characteristic (ROC) curve of 10-fold cross validations of each predictive modelling approaches.

Variables included in the prediction models

Table 4 presents the variables selected for each model. Due to the failure of the administrative data analysed in isolation to provide predictions above chance, variable selections for Model 1 are considered unreliable and are marked with a circle for comparison purposes.

There was considerable concordance in the variables deemed useful for prediction across models. Principally, variables representing clinical information, including usage of hospital services within the past 12 months (number of admissions, emergency department (ED) presentations, and outpatient clinic attendance), number of past medical histories recorded in the progress notes, social history, and language spoken other than English were selected across multiple models, including by the elastic net. Variables relating to written discharge documentation were also selected, including completion of nursing admission and discharge planning documentation and date of discharge summary issued. The elastic net did not select any administrative variable and the gradient boosted tree selected only one administrative variable in Model 3 (distance from the hospital).

Languages spoken other than English/interpreter services requirement was selected by the elastic net and stepwise logistic regression models. It is worth noting that a total of five patients in the dataset required interpreter service at the index admission and all of them experienced 30-day unplanned readmission. The low cell count potentially precludes the variable from emerging as a useful predictor in other models, and suggests caution in interpreting the influence of this variable given the low count.

Discussion

We present a matched case-control study using retrospective analysis of patients' medical records to identify paediatric 30-day

all-cause same-hospital unplanned readmissions. Model prediction improvements were identified when adding clinical information and written discharge documentation compared with the available administrative data. Previous paediatric studies^{3,11–14,33–35} reported predictive model performance, with only one study³⁴ examining both clinical and administrative data by reviewing patients' medical record charts. Previous studies that have applied machine learning to paediatric readmission prediction obtained similar¹³ or better performance^{11,12} to the current study. However, our study used a matched case-control design, with matching across age, sex, and diagnosis that may better identify factors contributing to readmission, distinct from diagnosis.

Four of the identified predictors in this study were consistent with previous research, including the number of hospitalisations prior to the index admission,²³ day of discharge,³⁶ LOS,³³ and the number of comorbidities.^{2,23,24,33,37} Previous studies have also investigated socioeconomic status in terms of using the area-level deprivation^{2,38} and type of health insurance.^{24,33,37,38} This study extracted patients' significant social history (e.g. under the care of the Department for Child Protection) from their medical records and found a positive association with readmissions. The use of an interpreter service was also selected as a predictor of unplanned readmissions; however, interpreter service usage was only selected by two of the machine learning models (stepwise logistic regression and elastic net), suggesting some caution when interpreting the utility of this variable in predicting readmission. Furthermore, there is inconsistency in the literature with respect to this variable and how well it is able to predict readmission. Previous studies^{6,39} that have examined whether speaking a language other than English was associated with unplanned hospital readmissions have been inconclusive due to low numbers of cases in the dataset,⁴⁰ as was the case in this study's dataset. Future studies could examine whether a

Table 4. Variables selected by each model
 GLM, logistic regression; G-S, stepwise logistic regression; RF, random forest; EN, elastic net; XGB, gradient boosted tree

Variables	Model 1					Model 2					Model 3					
	GLM	G-S	RF	EN	XGB	GLM	G-S	RF	EN	XGB	GLM	G-S	RF	EN	XGB	
Administrative	Age			○	○											
	Sex (male)			○												
	Insurance status (private)			○												
	Admission status (emergency)			○												
	Socioeconomic Indexes for Areas (SEIFA) percentile			○		○										
	Distance from residential address to hospital	○	○			○										X
	Length of stay at index admission			○		○			X					X		
	No. co-diagnosis			○		○			X					X		
	Had general anaesthetic at index admission			○		○								X		
	Source of referral transport (ambulance)			○		○								X		
	Day of discharge date		○			○	X	X	X			X	X			
	Day of admission date	○	○			○	X	X	X			X	X			
	Day of admission (weekday/weekend and public holiday)	○	○				X	X				X	X			
	Clinical	No. admissions in the previous 12 months					X	X	X	X	X	X	X	X	X	X
No. emergency department presentations in the previous 12 months						X	X	X	X	X	X	X	X	X	X	
No. outpatient clinic attendances in the previous 12 months								X		X			X	X	X	
No. past medical histories recorded in the progress notes						X	X	X	X	X	X	X	X	X	X	
Significant social history								X				X	X	X	X	
Language spoken other than English (interpreter service required)								X				X		X		
Significant vital signs								X								
Known allergies								X				X				
Written discharge documentation		Completeness of <i>Nursing Admission and Discharge Planning Form</i> , Discharge Planning section (incompleteness)										X	X	X	X	X
		Completeness of <i>Nursing Admission and Discharge Planning Form</i> , Admission section (incompleteness)											X	X	X	X
	Delay in issuing discharge summary												X	X	X	
	Progress note or <i>Clinical Pathway</i> documentation at discharge by nurses (not recorded)											X				
	Follow-up information documented by doctors (not recorded)											X				

sample enriched in people requiring interpreter services contributes significantly to readmission.

Social history and English language proficiency are routinely assessed at the time of admission, and this study highlights the need for early commencement of discharge planning for these patients.¹⁹ Patients identified as having significant social history at the time of admission require a designated hospital-based social worker to assess and provide social needs for the family/caregiver. The social worker should also collaborate with other healthcare providers to implement a discharge planning process that ensures continuity of care at home, post-discharge.⁴¹ Interpreter services should be available throughout hospitalisation for families/caregivers with language barriers, and are crucial at the time discharge information is delivered by doctors and nurses. The ‘teach-back process’ is also recommended to ensure families’/caregivers’ understanding of the discharge information.⁴²

The quality of written discharge documentation was examined in this study. Incomplete nursing admission and discharge planning documentation, and delay in issuing discharge summaries were associated with unplanned readmissions. Previous research is inconsistent in reporting the association between written discharge documentation and readmissions. One study⁵ found that not providing a written instructional discharge plan to caregivers of children with asthma resulted in a 1.55 times higher readmission rate. A second study⁶ reported that having discharge follow-up plans contributed to readmissions; however, this result was possibly due to the low rate of primary care providers follow-up plan documentation in the discharge summary. A third study⁴ examined the association between asthma patients who were given follow-up appointments and asthma patient readmissions, but the results were inconclusive. Completeness of discharge documentation may reflect on the level of

comprehensiveness of discharge information conveyed to families/caregivers.⁸ However, our study conducted limited research into what and how the discharge information is communicated between healthcare providers and families/caregivers. A clinical observational study is, therefore, required to explore communication practice at discharge. It is imperative to complete and distribute discharge summaries to the caregiver's/family's general practitioner prior to sending a patient home.^{7,43} Discharge summaries contain detailed admission information for when the patient seeks medical advice following hospital discharge, and therefore may prevent unnecessary return ED visits or even unplanned readmissions.

This is the first known study using machine learning approaches to predict paediatric unplanned readmissions in Australia. Stepwise logistic regression, random forest, elastic net, and gradient boosted tree approaches were utilised and compared with standard logistic regression analysis. We found modestly greater prediction accuracy using machine learning for the identification of unplanned readmissions, especially using gradient boosted trees. Similarly, an adult population study¹⁷ also found substantially improved prediction of unplanned hospital readmissions using machine learning.

A limitation of this study is that principal diagnosis of the index admission was not examined as a predictor because it was used to match cases and controls. This study is also limited by a specified local context of WA. In comparison to the literature, this study was based on 470 matched case-controls, a small sample size, due to the difficulty and cost of auditing patients' medical records. Therefore, use of electronic medical records is warranted to allow easy access not only to clinical information but also to written discharge documentation information. A larger sample size is also required to further leverage the benefit of machine learning approaches in the development of predictive models for unplanned paediatric readmissions, as we used a highly constrained approach to prevent overfitting. This retrospective cohort study used historical data from 2010 to 2014, which may reduce the relevance to current clinical practice. However, risk factors associated with paediatric unplanned hospital readmissions have remained stable over the last decade, based on our recently published systematic review,¹ indicating that the datasets used in this study provided relevant information regarding current readmission factors.

Conclusions

Adding clinical information and written discharge documentation demonstrated incremental improvements in prediction of paediatric unplanned hospital readmissions. Machine learning approaches, especially gradient boosted trees, achieved improved prediction accuracy over standard logistic regression analysis. Social and written discharge documentation variables including social history, poor English language proficiency, incomplete discharge documentation, and delay in issuing discharge summary, add value to prediction and our understanding of unplanned hospital readmissions. These predictors could also be translated into clinical practice of discharge planning to help prevent paediatric 30-day all-cause same-hospital unplanned readmission.

Competing interests

The authors have no competing interests to declare.

Acknowledgements

The authors acknowledge staff at the Health Information & Administrative Services, Child & Adolescent Health Service, for their assistance in retrieving medical records, especially Dr Julia Logan, Head of Department. This study was supported by a grant from the Australian Research Council – ARC Linkage Grant (Project ID: LP140100563) and the Chief Investigator is PRD. HZ was also supported by an Academic Support Grant 2016 and an Academic Research Grant 2014 from the Nursing and Midwifery Office, WA Department of Health.

References

- Zhou H, Roberts PA, Dhaliwal SA, Della PR. Risk factors associated with paediatric unplanned hospital readmissions: a systematic review. *BMJ Open* 2019; 9: e020554. doi:10.1136/bmjopen-2017-020554
- Wijlaars LP, Hardelid P, Woodman J, Allister J, Cheung R, Gilbert R. Who comes back with what: a retrospective database study on reasons for emergency readmission to hospital in children and young people in England. *Arch Dis Child* 2016; 101: 714–8. doi:10.1136/archdischild-2015-309290
- Minhas SV, Chow I, Feldman DS, Bosco J, Otsuka NY. A predictive risk index for 30-day readmissions following surgical treatment of pediatric scoliosis. *J Pediatr Orthop* 2016; 36: 187–92. doi:10.1097/BPO.0000000000000423
- Feng JY, Toomey SL, Zaslavsky AM, Nakamura MM, Schuster MA. Readmissions after pediatric mental health admissions. *Pediatrics* 2017; 140: e20171571. doi:10.1542/peds.2017–1571
- Topal E, Gucenmez OA, Harmanci K, Arga M, Derinoz O, Turktas I. Potential predictors of relapse after treatment of asthma exacerbations in children. *Ann Allergy Asthma Immunol* 2014; 112: 361–4. doi:10.1016/j.anai.2014.01.025
- Coller RJ, Klitzner TS, Lerner CF, Chung PJ. Predictors of 30-day readmission and association with primary care follow-up plans. *J Pediatr* 2013; 163: 1027–33. doi:10.1016/j.jpeds.2013.04.013
- Choudhry AJ, Baghdadi YMK, Wagie AE, Habermann EB, Cullinane DC, Zielinski MD. Readability of discharge summaries: with what level of information are we dismissing our patients? *Am J Surg* 2016; 211: 631–36. doi:10.1016/j.amjsurg.2015.12.005
- Coghlin DT, Leyenaar JK, Shen M, Bergert L, Engel R, Hershey D, Mallory L, Rassbach C, Woehrlen T, Cooperberg D. Pediatric discharge content: a multisite assessment of physician preferences and experiences. *Hosp Pediatr* 2014; 4: 9–15. doi:10.1542/hpeds.2013-0022
- Olsen MR, Hellzen O, Skotnes LH, Enmarker I. Content of nursing discharge notes: associations with patient and transfer characteristics. *Open Nurs J* 2012; 2: 277–87. doi:10.4236/ojn.2012.23042
- Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. *Comput Methods Programs Biomed* 2018; 164: 49–64. doi:10.1016/j.cmpb.2018.06.006
- Jovanovic M, Radovanovic S, Vukicevic M, Pouke SV, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artif Intell Med* 2016; 72: 12–21. doi:10.1016/j.artmed.2016.07.003
- Stiglic G, Wang F, Davey A, Obradovic Z. Pediatric readmission classification using stacked regularized logistic regression models. *AMIA Annual Symp Proc* 2014; 2014: 1072–81.
- Wolff P, Grana M, Rios SA, Yarza MB. Machine learning readmission risk modeling: a pediatric case study. *BioMed Res Int* 2019; 2019: 8532892. doi:10.1155/2019/8532892

- 14 Janjua MB, Reddy S, Samdani AF, Welch WC, Ozturk AK, Price AV, Weprin BE, Swift DM. Predictors of 90-day readmission in children undergoing spinal cord tumor surgery: a nationwide readmissions database analysis. *World Neurosurg* 2019; 127: e697–706. doi:10.1016/j.wneu.2019.03.245
- 15 Wiens J, Shenoy E. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis* 2018; 66: 149–53. doi:10.1093/cid/cix731
- 16 Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, Bhatt DL, Fonarow GC, Laskey WK. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017; 2: 204–209. doi:10.1001/jamacardio.2016.3956
- 17 Yang C, Delcher C, Shenkman E, Ranka S. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), 14–16 September 2016, Munich, Germany. IEEE; 2016. doi:10.1109/HealthCom.2016.7749452
- 18 Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, Hisamitsu T, Kojima G, Felsted J, Kakarmath S, Kvedar J, Jethwani K. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Med Inform Decis Mak* 2018; 18: 44. doi:10.1186/s12911-018-0620-z
- 19 Zhou H, Della P, Porter P, Roberts P. Risk factors associated with 30-day all-cause unplanned hospital readmissions at a tertiary children's hospital in Western Australia. *J Paediatr Child Health* 2020; 56: 524–46. doi:10.1111/jpc.14492
- 20 Child and Adolescent Health Service. History and design: Princess Margaret Hospital. Available at: <https://pch.health.wa.gov.au/About-us/History/Princess-Margaret-Hospital> [verified 16 February 2021].
- 21 Zhou H, Della P, Roberts P, Porter P, Dhaliwal S. A 5-year retrospective cohort study of unplanned readmissions in an Australian tertiary paediatric hospital. *Aust Health Rev* 2019; 43: 662–71. doi:10.1071/AH18123
- 22 Blackwell M, Iacus S, King GP. G. cem: coarsened exact matching in Stata. *Stata J* 2009; 9: 524–46. doi:10.1177/1536867X0900900402
- 23 Beck CE, Khambalia A, Parkin PC, Raina P, Macarthur C. Day of discharge and hospital readmission rates within 30 days in children: a population-based study. *Paediatr Child Health* 2006; 11: 409–12. doi:10.1093/pch/11.7.409
- 24 Berry J, Hall DE, Kuo DZ, Cohen E, Agrawal R, Feudtner C, Hall M, Kueser J, Kaplan W, Neff J. Hospital utilization and characteristics of patients experiencing recurrent readmissions within children's hospital. *JAMA* 2011; 305: 682–90. doi:10.1001/jama.2011.122
- 25 Schlesselman JJ, Stolley PD. Case-control studies: design, conduct, analysis. New York: Oxford University Press; 1982.
- 26 Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28: 112–8. doi:10.1093/bioinformatics/btr597
- 27 R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical computing; 2018. Available at: <https://www.R-project.org/> [verified 20 October 2020].
- 28 Kuhn M. The caret package. 2019. Available at: <https://cran.r-project.org/web/packages/caret/index.html> [verified 12 October 2020].
- 29 Tibshirani R. Regression shrinkage and selection via the Lasso. *JSTOR. Series B (Methodological)* 1996; 58: 267–88. Available at: <https://www.jstor.org/stable/2346178> [verified 12 October 2020].
- 30 Zou H, Hastie T. Regularization and variable selection via the elastic net. *JSTOR. Series B (Methodological)* 2005; 67: 301–20. Available at: <https://www.jstor.org/stable/3647580?seq=1> [verified 12 October 2020].
- 31 Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; 2/3: 18–22. Available at: <https://cogms.northwestern.edu/cbm/LiawAndWiener2002.pdf> [verified 12 October 2020].
- 32 Chen T, He T. xgboost: eXtreme gradient boosting. Package version 1.2.0.1. 2020. Available at: <https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf> [verified 1 May 2019].
- 33 Feudtner C, Levin JE, Srivastava R, Goodman DM, Slonim AD, Sharma V, Shah SS, Pati S, Fargason C, Hall M. How well can hospital readmission be predicted in a cohort of hospitalized children? A retrospective, multicenter study. *Pediatrics* 2009; 123: 286–93. doi:10.1542/peds.2007-3395
- 34 Sacks JH, Kelleman M, McCracken C, Glanville M, Oster M. Pediatric cardiac readmissions: an opportunity for quality improvement? *Congenit Heart Dis* 2017; 12: 282–8. doi:10.1111/chd.12436
- 35 Vo D, Zurakowski D, Faraoni D. Incidence and predictors of 30-day postoperative readmission in children. *Pediatric Anaesth* 2018; 28: 63–70. doi:10.1111/pan.13290
- 36 Auger K, Davis M. Pediatric weekend admission and increased unplanned readmission rates. *J Hosp Med* 2015; 10: 743–45. doi:10.1002/jhm.2426
- 37 Khan A, Nakamura MM, Zaslavsky AM, Jang J, Berry JG, Feng JY, Schuster MA. Same-hospital readmission rates as a measure of pediatric quality of care. *JAMA Pediatr* 2015; 169: 905–12. doi:10.1001/jamapediatrics.2015.1129
- 38 Sills MR, Hall M, Colvin JD, Macy ML, Cutler GJ, Bettenhausen JL, Morse RB, Auger KA, Raphael JL, Gottlieb LM, Fieldston ES, Shah SS. Association of social determinants with children's hospitals' preventable readmissions performance. *JAMA Pediatr* 2017; 170: 350–8. doi:10.1001/jamapediatrics.2015.4440
- 39 Richards MK, Yanez D, Goldin AB, Grieb T, Murphy WM, Drugas GT. Factors associated with 30-day unplanned pediatric surgical readmission. *Am J Surg* 2016; 212: 426–32. doi:10.1016/j.amjsurg.2015.12.012
- 40 Tommey S, Peltz A, Loren S, Tracy M, Williams K, Pengeroth L, Ste Marie A, Onorato S, Schuster MA. Potentially preventable 30-day hospital readmissions at a children's hospital. *Pediatr Neonatol* 2016; 138: e20154182. doi:10.1542/peds.2015-4182
- 41 Heenan D, Birrell D. Hospital-based social work: challenges at the interface between health and social care. *Br J Soc Work* 2019; 49: 1741–58. doi:10.1093/bjsw/bcy114
- 42 Kornburger CK, Gibson C, Sadowski S, Maletta K, Klingbeil C. Using 'teach-back' to promote a safe transition from hospital to home: an evidence-based approach to improving the discharge process. *J Pediatr Nurs* 2013; 28: 282–91. doi:10.1016/j.pedn.2012.10.007
- 43 Hoyer EH, Odonkor CA, Bhatia SN, Leung C, Deutschendorf A, Brotman DJ. Association between days to complete inpatient discharge summaries with all-payer hospital readmissions in Maryland. *J Hosp Med* 2016; 11: 393–400. doi:10.1002/jhm.2556