doi:10.1111/imj.14962

REVIEW

Machine learning in the prediction of medical inpatient length of stay

Stephen Bacchi,[1,2] Yiran Tan,[1,2] Luke Oakden-Rayner,[1,2] Jim Jannes,[1,2] Timothy Kleinig[1,2] and Simon Koblar[1,2]

[1]Royal Adelaide Hospital, and [2]Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, South Australia, Australia

Key words
artificial intelligence, deep learning, neural network, predictive analytics.

Correspondence
Stephen Bacchi, Neurology Department, Royal Adelaide Hospital, Port Road, Adelaide, SA 5000, Australia.
Email: stephen.bacchi@sa.gov.au

Received 30 November 2019; accepted 16 June 2020.

Abstract

Length of stay (LOS) estimates are important for patients, doctors and hospital administrators. However, making accurate estimates of LOS can be difficult for medical patients. This review was conducted with the aim of identifying and assessing previous studies on the application of machine learning to the prediction of total hospital inpatient LOS for medical patients. A review of machine learning in the prediction of total hospital LOS for medical inpatients was conducted using the databases PubMed, EMBASE and Web of Science. Of the 673 publications returned by the initial search, 21 articles met inclusion criteria. Of these articles the most commonly represented medical specialty was cardiology. Studies were also identified that had specifically evaluated machine learning LOS prediction in patients with diabetes and tuberculosis. The performance of the machine learning models in the identified studies varied significantly depending on factors including differing input datasets and different LOS thresholds and outcome metrics. Common methodological shortcomings included a lack of reporting of patient demographics and lack of reporting of clinical details of included patients. The variable performance reported by the studies identified in this review supports the need for further research of the utility of machine learning in the prediction of total inpatient LOS in medical patients. Future studies should follow and report a more standardised methodology to better assess performance and to allow replication and validation. In particular, prospective validation studies and studies assessing the clinical impact of such machine learning models would be beneficial.

Introduction

The accurate prediction of length of stay (LOS) in hospitals can aid in bed management and hospital staffing decisions.[1] However, LOS may be influenced by many factors, particularly in complex medical patients, and may be difficult to predict. Machine learning refers to the use of computers to discover patterns within data, without a human explicitly programming how to do so.[2] Given the assumption-free data-driven nature of machine learning it can be hypothesised that it may be able to assist in the accurate prediction of LOS for medical patients.

Many medical applications of machine learning involve making individual patient predictions. If the predictions place individuals into categories (such as predicting LOS as either ≥7 days or <7 days) then this is commonly referred to as a 'classification task'. Conversely, if a continuous outcome (e.g. prediction of LOS as the actual number of days that a patient will be in hospital) is predicted, it is generally referred to as a 'regression task'.[3] These types of study have different model performance metrics. Classification studies typically report a combination of prevalence-dependent performance metrics (such as accuracy, positive predictive value and negative predictive value) and prevalence-independent performance metrics (such as area under the receiver operator curve (AUC), sensitivity and specificity). There is ongoing discussion as to which outcome metrics are ideally presented in different instances;[4,5] however, a combination of metrics provides the most comprehensive representation of model performance. Regression studies typically present performance metrics as mean absolute error, mean squared error, root mean squared error and R^2.

Funding: None.
Conflict of interest: None.

Although there are a variety of conceptual frameworks, the development of a clinical machine learning application adopts a similar staged approach to the development of a clinical decision rule. For example, these stages typically involve a 'derivation' study, an 'external validation' study and then an 'impact/implementation' study.[6,7] In derivation studies for both classification and regression tasks, it is common for data from one population to be split into 'training' and 'testing' datasets. The training dataset is used for the development of the model. Performance is then assessed on the testing dataset (which is comprised of data from the same population that was separated for this purpose). In contrast, in an external validation study the performance of a previously derived model is assessed on a 'testing' dataset comprised of out-of-sample data, that is, data from a different clinical setting.

Awad *et al*. published a review regarding LOS prediction with machine learning (ML) in 2017.[8] However, this review focussed on explaining and summarising the methods of the reviewed studies, rather than critically appraising the studies. The critical appraisal of clinical machine learning research is an ongoing issue. While critical appraisal tools for predictive modelling derivation studies exist, such as the Checklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies[9] and Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement,[10] there are currently no critical appraisal tools with an explicit focus on machine learning. The TRIPOD-ML statement is currently in development.[11] It should be noted that critical appraisal of impact/implementation studies will require a different type of critical appraisal from that required for derivation and external validation studies. In accordance with these different requirements, other tools such as CONSORT-AI and SPIRIT-AI are currently in development,[12] expanding the existing CONSORT and SPIRIT statements on trial design to specifically address issues with ML.

This review was conducted with the aim of identifying previously published articles investigating the application of machine learning to the prediction of total hospital inpatient LOS for medical patients, critically appraising their methodology and evaluating the stage of development and implementation of such models.

Methods

This review was constructed according to the preferred reporting items for systematic review and meta-analysis protocols guidelines.[13] In September 2019, the databases PubMed, EMBASE and Web of Science were searched from their inception for articles relating to machine learning and LOS prediction in medical patients. The search terms (searched for in 'All Fields') were: ('Machine learning' OR 'artificial intelligence' OR 'deep learning' OR 'predictive analytics') AND ('length of stay' OR 'estimated discharge date' OR 'length of hospital stay') (see Supporting Information S1 for individual database search strings). The reference lists of included articles were then searched for further articles that fulfilled inclusion criteria.

Inclusion criteria were applied to the titles and abstracts of the articles returned by the search. If it could not be determined whether an article fulfilled the inclusion criteria, the article was retrieved in full text.

For inclusion in the review, a study was required to meet all of the following eligibility criteria:

1 Be published in English;
2 Be a primary research project (i.e. not a review or editorial);
3 Use machine learning for classification or regression (beyond that involved in regular medical statistical hypothesis testing) to predict LOS (see criteria 4);
4 Predict total inpatient LOS as an individual outcome and present performance metrics of this prediction relative to actual LOS (i.e. LOS prediction must be presented alone, and not solely as part of a composite end-point). LOS prediction for specific services during admission (e.g. LOS of time in intensive care unit (ICU), without total inpatient LOS), was not considered to fulfil this criterion;
5 Predict LOS for patients either specifically in an adult medical specialty, or for a group including patients from adult medical specialties (e.g. studies assessing all hospital inpatients were included, whereas studies specifically on surgical patients were excluded);
6 Be an article published in a peer-reviewed resource (abstracts from conferences and supplementary information were excluded);
7 Be available in full text to the authors conducting the review.

Quality analysis was conducted using a critical appraisal framework adapted from the TRIPOD statement[10,14] (Supporting Information S2). Data extraction was performed for the components of the quality analysis, in addition to the key results of each study (namely the outcome metrics of the best performing model in each instance). Eligibility determination was performed in duplicate in instances of borderline eligibility, and otherwise performed by a single author. Quality analysis and data extraction were conducted in duplicate using a standardised form. Instances of disagreement were resolved by discussion.

Results

The initial search returned 673 publications. Following the review of titles and abstracts, 570 publications were excluded (Fig. 1). One hundred and three articles were then reviewed in full text and their reference lists searched for further relevant studies, resulting in the inclusion of 21 articles in the review. Of these, 10 examined specific medical patient populations in the specialties of cardiology,[15–19] endocrinology (diabetes mellitus),[20] geriatrics,[21] infectious diseases (sepsis),[22] neurology (stroke)[23] and thoracic medicine (tuberculosis)[24] (Table 1). Eight studies included all inpatients at their respective centres, which encompassed medical patients[27–34] (Table 2). Three studies included patients with acute kidney injury (AKI),[35] ICU admissions[26] and elective admissions,[25] and met inclusion criteria due to the likely involvement of medical patients.

Models used in the located studies included support vector machines, artificial neural networks, Bayesian networks, decision tree algorithms, random forest algorithms and logistic regression models. Recurrent neural networks and convolutional neural networks were infrequently employed. The models were typically employed on data collected within the first 12–48 h of admission to make LOS predictions. However, there were also instances that used new data that became available throughout the course of the admission to make recurrent LOS predictions.[24] The majority of studies used combinations of demographic (e.g. age and gender), administrative (e.g. insurance status and whether admitted on weekend), clinical (e.g. vital signs, and comorbidities), laboratory (e.g. creatinine, haemoglobin, and bicarbonate) and treatment (e.g. prescribed medications) data to predict LOS. Types of data that were less frequently used to aid in LOS prediction included imaging data (e.g. radiology) and natural language data (e.g. from patient notes).

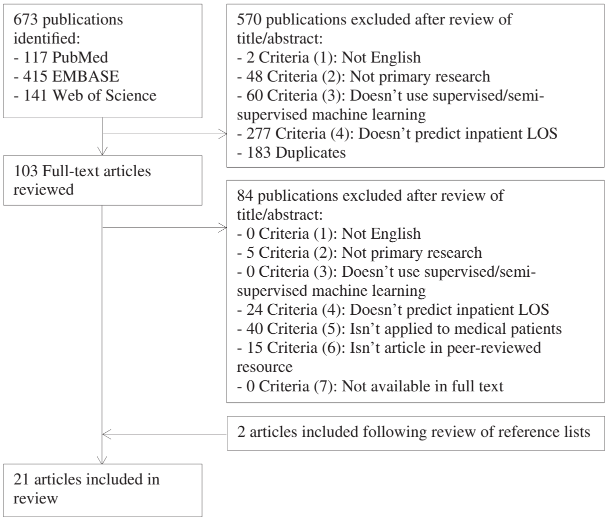
673 publications identified:
- 117 PubMed
- 415 EMBASE
- 141 Web of Science
570 publications excluded after review of title/abstract:
- 2 Criteria (1): Not English
- 48 Criteria (2): Not primary research
- 60 Criteria (3): Doesn't use supervised/semi-supervised machine learning
- 277 Criteria (4): Doesn't predict inpatient LOS
- 183 Duplicates
103 Full-text articles reviewed
84 publications excluded after review of title/abstract:
- 0 Criteria (1): Not English
- 5 Criteria (2): Not primary research
- 0 Criteria (3): Doesn't use supervised/semi-supervised machine learning
- 24 Criteria (4): Doesn't predict inpatient LOS
- 40 Criteria (5): Isn't applied to medical patients
- 15 Criteria (6): Isn't article in peer-reviewed resource
- 0 Criteria (7): Not available in full text
2 articles included following review of reference lists
21 articles included in review

Figure 1 Diagram demonstrating the results from a search and application of eligibility criteria to identify articles that have used machine learning to predict the total inpatient length of stay for medical admissions.

Many of the studies lacked a detailed description of study design elements according to the criteria in the employed critical appraisal framework. In particular, many studies did not provide clear inclusion criteria for the patients in the study (5/21), demographic details for the included patients (12/21) or details regarding the frequency of medical conditions/comorbidities for the included patients (13/21). Studies infrequently defined a primary objective or reported the number of patients screened for inclusion (9/21).

Specifically assessing the machine learning methodology, studies often did not specify their approach to handling missing data (11/21), and were often unclear in their description of the training/testing methodology employed. Seven studies appeared to use the same dataset for testing their models as they did for model development, without specifying that hold-out test data was employed (e.g. using k-fold cross-validation over the entire dataset to derive performance metrics, without specifying the use of hold-out test data in each fold). There were also multiple studies that did not provide the proportion or distribution of the LOS in the test set being evaluated.

All but one of the identified studies used retrospective datasets,[21] and none of the identified studies prospectively externally validated previously derived models in new datasets. Furthermore, none of the identified studies evaluated the impact of the real-world implementation of their LOS prediction models.

Studies focussing on medical specialty patients

Cardiology was the most frequently studied medical specialty. Of the five studies in this area, two focussed on multiple-cause cardiology admissions,[15,16] one study focussed on patients with heart failure,[17] one focussed on patients with coronary artery disease,[18] and one focussed on patients with unstable angina.[19] One of the most clearly written of these studies examined all-cause cardiology admission LOS prediction with a variety of models in 16 414 admissions from a hospital in Saudi

Table 1 Studies predicting length of stay (LOS) of medical inpatients from individual specialties

Reference	Specialty	Retrospective vs prospective	Eligibility criteria	Sample size	Models used	LOS proportion or distribution	Regression or classification outcome	If LOS classification, thresholds employed	Model performance	Critical appraisal
Tsai *et al*. 2016[15]	Cardiology	Retrospective	Coronary atherosclerosis, heart failure or acute myocardial infarction	2377	Logistic regression and artificial neural network	Graph presents LOS distribution	Both	2-day tolerance	Accuracy AMI/CHF: 63.7%–65.7%; CAS: 88.3%–89.7%; AMI/CHF: MAE 3.87–3.97; CAS: 1.03–1.07; AMI/CHF: MRE 0.73–0.77; CAS MRE: 0.44–0.47	Clearly specified train/test split. Uncertain approach to missing data
Daghistani *et al*. 2019[16]	Cardiology	Retrospective	All adult cardiology admissions	16 414	Random forest, artificial neural network, support vector machine and Bayesian network	<3 days = 5063. 3–5 days = 5490. >5 days = 5861	Classification	<3 days, 3–5 days and > 5 days	Accuracy 80%; PPV 80%; sensitivity 80% AUC 0.94; RMSE 0.31; *F* score 80%	Clearly described LOS proportions. No ethics statement included
Turgeman *et al*. 2017[17]	Cardiology: CCF	Retrospective	All patients with admissions who had been diagnosed with CHF (although admission could be any cause)	20 321	Regression tree (Cubist)	Mean LOS 6.24, median 4, standard deviation 8.475	Regression	NA	MAE 1; R^2 0.79	Few details on patient medical conditions. Clearly described approach to missing data
Hachesu *et al*. 2013[18]	Cardiology: IHD	Retrospective	All had coronary artery disease	2064	ANN, SVM and decision tree	LOS (days): 0–5, 35.8%; 6–9, 24.9%; and ≥10, 39.3%	Classification	LOS 0–5 days, 6–9 days and ≥10 days	96.4% accuracy; 97.3% sensitivity; 98.1% specificity	Included patient demographic and comorbidity details. No ethics statement included
Huang *et al*. 2016[19]	Cardiology: unstable angina	Retrospective	All unstable angina admissions	3492	Multiple models including treatment pattern models and multi-label k-nearest neighbours	Describes LOS typically being between 2 and 3 weeks	Classification	LOS ≤7 days, 8–14 days, 14–28 days, >28 days	Accuracy 0.849	Included patient demographic and comorbidity details. Uncertain how many individuals were screened for inclusion
Morton *et al*. 2014[20]	Endocrinology	Retrospective	Not specified	10 000	Multiple models including random forest and multiple linear regression	Uncertain	Classification	LOS <3 days or ≥3 days	Accuracy 0.68 (±0.01); AUC 0.76 ± 0.01	Uncertain proportion/distribution of LOS. Reported prevalence dependent and independent performance
Launay *et al*. 2015[21]	Geriatrics	Prospective	Age ≥80 years	993	Artificial neural network (multi-layer perceptrons)	LOS ≥13: 21.6% in training set, and 24.9% in test set	Classification	LOS ≥13 days or <13 days	Accuracy 87.4; AUC 90.5; specificity 96.6%; sensitivity 62.7%; PPV 87.1; NPV 87.5	Clearly described LOS proportions in train/test sets. Clearly described train/test split

Table 1 *Continued*

Reference	Specialty	Retrospective vs prospective	Eligibility criteria	Sample size	Models used	LOS proportion or distribution	Regression or classification outcome	If LOS classification, thresholds employed	Model performance	Critical appraisal
Tsoukalas *et al.* 2015[22]	ICU: Sepsis	Retrospective	≥18 years of age, ICU admission, meeting ≥2 SIRS criteria	1492	Support vector machine	Mean LOS 17.0 (standard deviation 36.7 days)	Classification	4, 8 and 12 days	Accuracy 0.69–0.82; AUC 0.69–0.73	Reported prevalence dependent and independent performance. Discussion of improved outcomes is unclear
Al Taleb *et al.* 2017[23]	Neurology: stroke	Uncertain	Not specified	716	Decision tree algorithm and Bayesian network	Uncertain	Classification	LOS 0–2 days, 3–7 days, 8–16 days and >16 days	Accuracy 81.29%; AUC 0.936; sensitivity 0.813; specificity 0.896	Uncertain proportion/ distribution of LOS. Uncertain inclusion criteria
Huang *et al.* 2013[24]	Respiratory infections	Retrospective	Admissions with a primary diagnosis ICD code consistent with tuberculosis	284	Temporal similarity model	Mean LOS 13.6	Regression	NA	RMSE variable depending on how many days of data into the admission the patient was (from approximately 8–1.75)	Distinctive approach of making repeated predictions of LOS during admission. Comparatively small sample size

AMI, acute myocardial infarction; ANN, artificial neural network; AUC, area under the receiver operator curve; CAS, coronary atherosclerosis; CCF, congestive cardiac failure; CHF, congestive heart failure; ICU, intensive care unit; IHD, ischaemic heart disease; MAE, mean absolute error; MRE, mean relative error; NPV, negative predictive value; PPV, positive predictive value; RMSE, root mean squared error; SIRS, systemic inflammatory response syndrome; SVM, support vector machine.

Table 2 Studies predicting length of stay (LOS) of groups of inpatients that included medical patients

Citation	Specialty	Retrospective vs prospective	Eligibility criteria	Sample size	Models used	LOS proportion or distribution	Regression or classification outcome	If LOS classification, thresholds employed	Model performance	Critical appraisal
Steele and Thompson 2019[25]	All elective admissions	Retrospective	Not specified	242 024	Multiple models including Naïve Bayes and k-nearest neighbours	Uncertain	Classification	≥8 days or < 8 days	AUC 0.904; specificity 0.92; AUCPR: 0.933; FN rate: 0.331	Large sample size. Few details on patient medical conditions
Sotoodeh and Ho 2019[26]	All ICU admissions	Retrospective	Existing dataset	4000	Hidden Markov model	Uncertain	Regression	NA	RMSE 228.12	Clearly described method for management of missing data. No ethics statement
Stojanovic *et al*. 2017[27]	All inpatient admissions	Retrospective	Not specified	100 000	disease +procedures2vec	Total dataset mean LOS 3.71–5.94	Regression	NA	R^2 0.0766–0.4356	Diverse patient population. Limited discussion
Caetano *et al*. 2014[28]	All inpatient admissions	Retrospective	Not specified	26 431	Random forest	Uncertain	Regression	NA	R^2 0.813; MAE 0.224; RMSE 0.469	Uncertain LOS distribution. Specified approach to missing data
Livieris *et al*. 2018[29]	All inpatient admissions	Retrospective	Limited to age >65 years	2702	Two-level classifier using random forest and k-nearest neighbours	Majority of cases were 1-day stays, followed by ≥5 day stays	Classification	1, 2, 3, 4 or ≥5 days	Accuracy 78.5%	Clearly presented confusion matrix. Few details on patient medical conditions
Livieris *et al*. 2018[30]	All inpatient admissions	Retrospective	Limited to age >65 years	4403	A variety of semi-supervised learning models were used including naive Bayes and multi-layer perceptron	Uncertain	Classification	1–2, 3–6, >6 days	Accuracy 63.23%–65.30%	Few details on patient medical conditions Uncertain approach to missing data
Baek *et al*. 2018[31]	All inpatient admissions	Retrospective	All admissions	45 546	Regression and random forest models	Mean LOS 7.0 (IQR 2–8)	Both	≥30 days or <30 days	Accuracy 0.9732; MAE 4.68	Clearly described number of individuals screened for inclusion. Clearly described

Table 2 *Continued*

Citation	Specialty	Retrospective vs prospective	Eligibility criteria	Sample size	Models used	LOS proportion or distribution	Regression or classification outcome	If LOS classification, thresholds employed	Model performance	Critical appraisal
										approach to missing data
Cui *et al*. 2018[32]	All inpatient admissions	Retrospective	All admission except rare diagnoses	750 000	Multiple models including random forest, decision tree and neural network	Uncertain	Regression	NA	R^2 0.554; RMSE 3.10; MAE 2.19	Large sample size. Few details on patient medical conditions
Liu *et al*. 2010[33]	All inpatient admissions	Retrospective	Age ≥15 years and not hospitalised for childbirth	155 474	Logistic regression	Mean LOS 4.5 days ± 7.7	Regression	NA	R^2 0.146; MSE/1000 29.0	Discussed exclusion of individuals with incomplete data. Included demographic details of patients
Rajkomar *et al*. 2018[34]	All inpatients	Retrospective	Age ≥18 years and ≥24 h hospital admission	216 221	Recurrent neural networks	22.3%–24.2% long stays in different datasets	Classification	≥7 days or < 7 days	AUC 0.85–0.86	Included patient demographic and medical characteristics. Clearly described train/test methodology
Saly *et al*. 2017[35]	Medicine: all patients in a trial with AKI	Retrospective	Patients enrolled in AKI trial. Eligibility criteria as per AKI trial	2241	Random forest and logistic regression	Median LOS for whole cohort was 10.2 (6.0–17.2) days	Regression	NA	R^2 0.2 (0.14–026)	Included details on patient medical conditions. Discussed number of individuals screened for inclusion

AKI, acute kidney injury; AUC, area under the receiver operator curve; AUCPR, area under the precision-recall curve; FN, false negative rate; ICU, intensive care unit; IQR, interquartile range; MAE, mean absolute error; MSE, mean squared error; RMSE, root mean squared error.

Arabia.[16] This study employed a classification approach (<3 days, 3–5 days and >5 days) and with a random forest model found an AUC of 0.94, sensitivity 80% and accuracy 80%. Strengths of this study are that it included the proportions of the different classes of LOS in the study population, as well as demographic and medical details of the included patients. However, the study did not explain how it managed missing data and did not adequately describe the cross-validation procedure employed for model selection and assessment.

Of the other studies examining areas of interest to individual medical specialties, the strongest was a study of 993 geriatric patients.[21] Aspects of this study that made it of high quality included the definition of inclusion criteria (admission following visit to emergency department and age >80 years), prospective data collection, provision of demographic/medical details of included patients, and presenting details regarding the proportion of different outcome classes in the training and test sets (LOS ≥13 days accounted for 21.6% of training set, and LOS ≥13 days 24.9% in test set). This study also presented a range of prevalence-dependent performance metrics (accuracy 87.4%, positive predictive value 87.1%, negative predictive value 87.5%) and prevalence-independent performance metrics (AUC 0.905, specificity 96.6%, sensitivity 62.7%), as well as raw true/false positive/negative results, enabling the calculation of other metrics if required.[21]

The study by Huang *et al*. predicting LOS for patients with tuberculosis was also notable, given it used a different method for LOS prediction as compared to the other included studies.[24] While most other studies used data from a defined period at the start of an admission to predict LOS (typically 12–48 h), this study used ongoing data collection throughout a hospital admission to recurrently generate new LOS estimates. Although this study had a small sample size ($n = 284$), it demonstrated ongoing improvement in LOS prediction throughout the course of the hospital stay as more data became available.

Studies encompassing all inpatient admissions, including medical patients

Eight studies examined the prediction of total inpatient LOS in all patients admitted to given centres, including medical patients. Three studies examined all inpatient admissions, including medical patients, that met a clinical criterion, namely ICU admission, AKI or elective admission.[25,26,35] Of these studies, the highest quality in terms of reporting was conducted by Rajkomar *et al*. This study used retrospective datasets from two hospitals in the USA in all ≥24 h inpatient admissions in ≥18-year-old patients ($n = 216\,221$) to derive classification models predicting LOS ≥7 days or <7 days.[34] This study was notable because of inclusion of patient demographics/medical information, as well as clear descriptions of the machine learning methodologies employed, and details on the proportions of the LOS classes in different datasets (LOS ≥7 days 22.3%–24.2%). This study found an AUC of 0.85–0.86 in this LOS classification.[34]

Three other studies also assessed models predicting all inpatient LOS as a classification task.[29–31] These studies reported accuracies of 78.5%,[29] 63.2%–65.3%[30] and 97.3%.[31] Studies that evaluated LOS as a regression task reported mean absolute errors including 0.224,[28] 2.19[32] and 4.68.[31] However, it is difficult to compare results among the identified studies for a variety of reasons. These reasons include that different studies employed different classification thresholds (e.g. predicting ≥30 days vs <30 days or predicting ≥7 days vs <7 days), approached LOS prediction as a regression task or classification task variably, presented different outcome metrics and had differing datasets (Table 2).

Discussion

The use of machine learning to predict total in-hospital LOS for medical patients has been assessed by several studies with variable methodologies and results. The wide range of model performances reported when similar models are applied to similar tasks likely reflects differences in methodology, or in patient population characteristics, between studies that have often not been described in sufficient detail. Common methodological issues include a lack of patient demographic/clinical information, failure to define a primary objective and failure to report the number of patients screened prior to inclusion. No studies performing prospective external validation or assessment of the implementation of machine learning models for LOS prediction were identified.

Aspects of ML methodology that could be improved frequently related to the use of and reporting regarding test datasets. Multiple studies appeared to use the same dataset for testing their models as they did for model development, without specifying that hold-out test data was employed. While cross-validation may be used as a means of reducing sampling error, this method can lead to over-fitting if applied improperly. It must be specified that model selection and model evaluation processes involve different data, even within folds.[36] The proportion or distribution of the variable being predicted (LOS) should be reported in test datasets, in addition to training datasets, as this information may be important when interpreting performance metrics.

The frequent methodological shortcomings identified in the included studies may at least in part reflect a difference in writing styles and target audiences between computer science researchers and medical researchers. In articles focussing on the development of new methods to apply to LOS prediction, there were generally fewer details regarding patients. In studies focussed more on the application of machine learning methods to new patient groups, more detail on patient factors was typically included. We believe that, regardless of the field of the target audience, it is necessary to provide patient demographic and disease prevalence details to be able to evaluate how to generalise the findings of a study to the patient population at another centre and to compare performance between studies.

The implications for future ML research in this area relate to standards of reporting and methods of analysis. ML studies reporting on the prediction of clinical outcomes (such as LOS) are required to present clear inclusion criteria and relevant clinical and demographic details of included patients, in order to enable clinicians at other centres to evaluate the possible external generalisability of the findings presented in the research. The proportion and distribution of outcomes of interest are required to be presented for test datasets in order to enable the interpretation of certain performance metrics. Future ML research in LOS may be able to utilise data types not frequently investigated in the identified studies, such as imaging data and natural language data. The generation of recurrent LOS predictions, using additional data accumulated throughout the admission, as opposed to data from only the first 12–48 h, may also be an area to investigate to improve performance.

In terms of current clinical practice, this review has shown that ML medical inpatient LOS prediction is a promising area, but that further research is required to support the use of such models. Currently there are no published studies reporting on prospective external validation of models to predict LOS in this cohort of patients. Similarly, no studies were identified that have implemented such models and demonstrated a benefit to patient or healthcare-system-oriented outcomes.

Limitations

The exclusion of non-English articles is a limitation of this review. It is a limitation that some studies had their eligibility for inclusion in the review determined by a single author. Publication bias may have influenced the results of the review. As discussed previously, it is difficult to compare the performance of models between studies, due to the differences in study design, outcome metrics and patient populations.

Conclusion

The variable performance reported by the studies identified in this review supports the need for further research on the utility of ML in the prediction of total inpatient LOS in medical patients. In particular, prospective external validation and implementation studies are required. Clinical machine learning external validation studies should aim to include clear definitions of which data are used for model development and testing, and the proportion/distribution of the outcome of interest in the testing set. Future research in this area should take note of the shortcomings identified in the studies performed to date. In particular, subsequent studies should include relevant clinical details to enable the assessment of generalisability of findings to other patient cohorts.

References

1 Robinson G, Davis L, Leifer R. Prediction of hospital length of stay. *Health Serv Res* 1966; **1**: 287–300.

2 Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med* 2018; **284**: 603–19.

3 Deo R. Machine learning in medicine. *Circulation* 2015; **123**: 1920–30.

4 Pinker E. Reporting accuracy of rare event classifiers. *NPJ Digit Med* 2018; **1**: 56.

5 Rajkomar A, Dai AM, Sun M, Hardt M, Chen K, Rough K *et al*. Reply: metrics to assess machine learning models. *NPJ Digit Med* 2018; **1**: 57.

6 Stiell IG, Bennett C. Implementation of clinical decision rules in the emergency department. *Acad Emerg Med* 2007; **14**: 955–9.

7 Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med* 1999; **33**: 437–47.

8 Awad A, Bader-El-Den M, McNicholas J. Patient length of stay and mortality prediction: a survey. *Health Serv Manage Res* 2017; **30**: 105–20.

9 Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG *et al*. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11**: e1001744.

10 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur J Clin Invest* 2015; **45**: 204–14.

11 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; **393**: 1577–9.

12 Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the

CONSORT and SPIRIT statements. *Lancet* 2019; **394**: 1225.

13 Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015; **350**: g7647.

14 Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–73.

15 Tsai PF, Chen PC, Chen YY, Song HY, Lin HM, Lin FM *et al*. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J Healthc Eng* 2016; **2016**: 1–11.

16 Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *Int J Cardiol* 2019; **288**: 140–7.

17 Turgeman L, May JH, Sciulli R. Insights from a machine learning model for predicting the hospital length of stay (LOS) at the time of admission. *Expert Syst Appl* 2017; **78**: 376–85.

18 Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc Inform Res* 2013; **19**: 121–9.

19 Huang Z, Dong W, Ji L, Duan H. Outcome prediction in clinical treatment processes. *J Med Syst* 2015; **40**: 1–13.

20 Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. 2014 *13th International Conference on Machine Learning and Applications*. 2014; 428–31.

21 Launay CP, Rivière H, Kabeshova A, Beauchet O. Predicting prolonged length of hospital stay in older emergency department users: use of a novel analysis method, the artificial neural network. *Eur J Intern Med* 2015; **26**: 478–82.

22 Tsoukalas A, Albertson T, Tagkopoulos I. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Med Inform* 2015; **3**: e11.

23 Al Taleb A, Hasanat M, Kahn M. Application of data mining techniques to predict length of stay of stroke patients. *International Conference on Informatics, Health and Technology (ICIHT)*. 2017.

24 Huang Z, Juarez JM, Duan H, Li H. Length of stay prediction for clinical treatment process using temporal similarity. *Expert Syst Appl* 2013; **40**: 6330–9.

25 Steele R, Thompson B. Data mining for generalizable pre-admission prediction of elective length of stay. *9th IEEE Annual Computing and Communication Workshop and Conference (CCWC)*. 2019.

26 Sotoodeh M, Ho J. Improving length of stay prediction using a hidden Markov model. *AMIA Jt Summits Transl Sci Proc* 2019; **6**: 425–34.

27 Stojanovic J, Gligorijevic D, Radosavljevic V, Djuric N, Grbovic M, Obradovic Z. Modeling healthcare quality via compact representations of electronic health records. *IEEE/ACM Trans Comput Biol Bioinform* 2017; **14**: 545–54.

28 Caetano N, Cortez P, Laureano R. Using data mining for prediction of hospital length of stay: an application of the CRISP-DM methodology. *16th International Conference on Enterprise Information Systems (ICEIS)*. 2014.

29 Livieris I, Kotsilieris T, Dimopoulos I, Pintelas P. Decision support software for forecasting patient's length of stay. *Algorithms* 2018; **11**: 199.

30 Livieris IE, Dimopoulos IF, Kotsilieris T, Pintelas P. Predicting length of stay in hospitalized patients using SSL algorithms. *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion - DSAI 2018*. 2018; 16–22.

31 Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. *PLoS One* 2018; **13**: e0195901.

32 Cui L, Xie X, Shen Z, Lu R, Wang H. Prediction of the healthcare resource utilization using multi-output regression models. *IISE Trans Healthc Syst Eng* 2019; **8**: 291–302.

33 Liu V, Kipnis P, Gould M, Escobar G. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med Care* 2010; **48**: 739–44.

34 Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M *et al*. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; **1**: 18.

35 Saly D, Yang A, Triebwasser C, Oh J, Sun Q, Testani J *et al*. Approaches to predicting outcomes in patients with acute kidney injury. *PLoS One.* 2017; **12**: e0169305.

36 Krstajic D, Buturovic L, Leahy D, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Chem* 2014; **6**: 1–15.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Supplementary Information S1. Individual database search strings.
Supplementary Information S2. Critical appraisal framework.