Deep Learning in the Prediction of Clinically Significant Outcomes in Stroke and General Medicine Patients

PhD thesis

To obtain the degree of Doctor of Philosophy (PhD) in Medicine at the University of Adelaide

By

Dr Stephen Daniel Bacchi

ID: 1606819

Supervisors:

Professor Simon Koblar

Professor Timothy Kleinig

Professor Jim Jannes

March 2022 Adelaide, Australia

Table of Contents

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Dr Stephen Bacchi

Abstract

Background

The need for novel strategies to improve outcome prediction and the categorisation of unstructured medical data will increase as the demands on hospitals, associated with the increasing age and complexity of admitted patients, continues to rise. Stroke is a highly specialised field, in which key performance indicators and discharge planning have an important role. General medicine is a field that encompasses a wide variety of multisystem and undifferentiated illnesses. It is possible that machine learning, in particular deep learning, may be able to assist with the prediction of clinically significant outcomes both in areas with highly specialised assessment and treatment considerations (such as stroke), as well as fields with a diverse mix of medical conditions and comorbidities (such as general medicine).

Method

This thesis comprised of studies using machine learning to predict clinically significant outcomes in stroke and general medicine inpatients. Initially a systematic review was conducted to evaluate the existing literature regarding the prediction of one such clinically significant outcome, length of stay, in medical inpatients. Derivation and validation studies were conducted to develop models for stroke inpatients to aid with the prediction of discharge independence, survival to discharge, discharge destination and length of stay. Stroke key performance indicator-automated extraction and clinical coding categorisation were undertaken in studies employing techniques including natural language processing. Natural language processing was applied to general medicine free-text data in pilot, derivation, and validation studies in the prediction of outcomes including discharge timing.

Results

The systematic review identified a particular lack of prospective validation studies for machine learning models developed to aid with length of stay prediction in medical inpatients. The stroke model derivation, prospective and external validation studies demonstrated the successful use of machine learning models in the prediction of outcomes relevant to discharge planning for stroke patients. For example, an area under the receiver operator curve (AUC) of 0.85 and 0.87 was achieved for the prediction of independence at the time of discharge in the prospective and external validation datasets respectively. The automated collection of stroke key performance indicators and the application of natural language processing to stroke clinical coding also demonstrated performance as high as an AUC of 0.95-1.00 in key performance indicator classification tasks. The general medicine pilot, derivation, prospective and external validation studies demonstrated the development and success of artificial neural networks in the prediction of discharge within the next 48 hours (AUC 0.78 and 0.74 in the prospective and external validation datasets respectively).

Conclusions

Machine learning models (including deep learning) can successfully predict clinically significant outcomes in stroke and general medicine patients.

Acknowledgements

I would to thank my supervisors, Prof Koblar, Prof Kleinig and Prof Jannes, for their support and guidance.

I would also like to thank my friends and collaborators for their feedback and perseverance: Dr Gluck, Dr Gilbert, Dr Chim, Dr Cheng, Dr Oakden-Rayner, Prof Menon, Dr Moey and Dr Tan.

Thank you to Carol and Sam, my parents.

Publications

Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T & Koblar S 2020, 'Machine Learning in the Prediction of Medical Inpatient Length of Stay', *Internal Medicine Journal*, https://doi.org/10.1111/imj.14962

Bacchi S, Oakden-Rayner L, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Stroke Prognostication for Discharge Planning with Machine Learning: A Derivation Study', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2020.07.046

Bacchi S, Oakden-Rayner L, Menon D, Moey A, Jannes J, Kleinig T & Koblar S 2021, 'Prospective and external validation of stroke discharge planning machine learning models', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.12.031.

Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, 'Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study, *Internal Medicine Journal*, https://doi.org/10.1111/imj.15678.

Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, 'Improving the accuracy of stroke clinical coding with open-source software and natural language processing', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.10.024.

Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study', *Intern Emerg Med*, https://doi.org/10.1007/s11739-019-02265-3.

Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Jannes J, Kleinig T & Koblar S 2021, ‘Mixed-data Deep Learning in Repeated Predictions of General Medicine Length of Stay: A Derivation Study’, *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02697-w.

Bacchi S, Gilbert T, Gluck S, Cheng J, Tan Y, Chim I, Jannes J, Kleinig T & Koblar S 2021, ‘Daily estimates of individual discharge likelihood with deep learning natural language processing in general medicine: a prospective and external validation study’, *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02816-7.

Chapter 1 - General Introduction

Demands on hospitals are increasing and novel strategies to aid with prognostication and service planning are required

The demands on hospitals are progressively increasing. For example, the Australian Institute of Health and Welfare has previously recorded that same day admissions increased at an average rate of 3.4% per year between 2014 and 2019 [1]. While the outbreak of COVID-19 temporarily reduced the number of hospitalisations, COVID-19 related disruptions will compound the challenges already faced by healthcare systems in future [2]. Accordingly, strategies to help accurately record hospital activity (to enable effective future planning and funding) and improve hospital efficiency (such as through effective inpatient prognostication and discharge planning) will continue to become increasingly relevant.

The increasing demands on hospital systems are multifactorial and have the potential to worsen hospital performance [3]. In Western countries such as Australia, this increasing demand may be due in part to an aging population, with multiple complex comorbidities [4]. For example, in recent Australian Institute of Health and Welfare statistics individuals over the age of 65 have accounted for approximately half of total days admitted to hospital, despite comprising less than one fifth of the population [5]. The increasing prevalence of chronic conditions in Western societies contributes to this complexity and demand [6]. In the setting of this increasingly elderly and comorbid population, COVID-19 has also presented unique challenges with respect to hospital demand. The demands placed on the healthcare system by COVID-19 affect multiple domains, including capacity planning with respect to available inpatient beds and workforce planning [3]. The implementation of new treatment strategies, such as novel stroke treatments, also require infrastructure and workforce planning that place additional demands on hospital resources. Collectively, these and other factors combine to explain why demands on hospital systems are increasing both in terms of the logistics of service provision as well as the associated fiscal burden, and highlight the importance of continuing to develop strategies to improve efficiency [3].

In addition to increasing demand, the increasing complexity of elderly and comorbid patients presents new challenges with respect to medical decision making and, in particular, the application of evidence-based medicine [7]. Evidence-based medicine, the cornerstone of current medical practice, involves the application of evidence gained through empiric study to populations to which the findings are thought to be generalisable. However, the majority of such studies are conducted in carefully selected populations with specific inclusion and exclusion criteria. Such criteria frequently result in individuals with advanced age and multiple comorbidities being excluded from these studies [8]. The exclusion of these patients means that the generalisability of certain treatments and prognostic scores to this population may be questionable [9]. While prognostic scores have been developed for a myriad of conditions, modern prognostic strategies will need to be dynamic and account for the interplay between multiple complex comorbidities in a given individual.

Accurate prognostication is important for patient counselling, medical decision making and service provision planning [10, 11]. The accurate prognostication of inpatient outcomes may facilitate the planning of aspects of hospital discharge [12]. Historically, prognostic scores typically comprised of multiple discrete data fields relating to an individual condition or a limited selection of comorbidities. One of the well-known prognostic scores, the Charlson Comorbidity Index, is comprised of 17-individual parameters [13]. A difficulty with such scores is that, by increasing the number of parameters to account for increased complexity, the scores may become unwieldy and impractical to use due to the time required to calculate the scores. One strategy to mitigate this issue is to develop new prognostic systems that use unstructured data that is readily available in electronic medical records. An alternative or complementary strategy would be to develop novel approaches to automatically collect and process unstructured data into a structured format that could then be used to calculate existing prognostic scores.

Automated data collection and processing has multiple medical applications, in addition to its possible use in prognostic scores, such as aiding with the monitoring of key performance indicators and the recording of casemix and hospital activity data. The monitoring of domain-specific key performance indicators is a high priority in multiple medical specialties to ensure high-quality and standardised care across diverse facilities [14]. However, the recording of such key performance indicators requires significant time and resources in an already strained system. Similarly, the recording of casemix information and hospital activity data is required for service provision planning and the allocation of activity-based funding, but requires significant time investments by specifically trained clinical coders in the recording and categorising of unstructured medical data [15]. The development of automatic means for the collection and categorisation of unstructured medical data for these tasks may improve efficiency and accuracy, and thereby likely improve aspects of hospital care planning and provision.

In a healthcare system with increasing demands and increasingly complex patients, novel strategies for the provision of accurate prognostic information and the collection and categorisation of unstructured medical data (such as for the monitoring of key performance indicators and activity levels) may enable improvements in efficiency and healthcare outcomes. These methods should be developed to be applicable both to fields with highly specialised assessment and treatment considerations (such as stroke), as well as fields with a diverse mix of medical conditions and comorbidities (such as general medicine).

Stroke is a specialised field in which key performance indicators and prognostication are important

Stroke is a leading cause of death and disability both in Australia and globally [16]. In 2019-2020 the Australian Bureau of Statistics described that cerebrovascular disease was the third most common cause of death in Australia [17]. However, stroke, in particular ischaemic stroke, also has time-critical interventions including thrombolysis and endovascular thrombectomy that may prevent significant disability and prove life-saving for appropriately selected patients [18]. Given that the efficacy of these interventions is highly time-dependent and may be associated with significant costs, this area highlights the importance of the accurate recording of key performance indicators and hospital activity levels. Despite these interventions (due to ineligibility or incomplete efficacy), for many individuals with stroke, rehabilitation remains a vital component of their recovery process [19]. Accurate prognostication to enable the selection of appropriate rehabilitation destinations is an important part of hospital resource planning that may facilitate a streamlined transition from acute to post-acute stroke care [20].

Key performance indicators in stroke may relate to acute stroke interventions (such as time from presentation to endovascular thrombectomy), prevention (such as the use of anticoagulants in atrial fibrillation) and rehabilitation (such as the time to be seen by a physiotherapist) [21]. The current Australian stroke key performance indicators are outlined in the Australian Commission on Safety and Quality in Health Care Acute Stroke Clinical

Care Standard and includes 17 unique indicators [22]. Previous studies have demonstrated that adherence to stroke key performance indicators is associated with reduced morbidity and mortality after stroke [23].

Previous prognostic scores for acute stroke have focussed on the prediction of outpatient outcomes, and rarely have been applied to the prediction of length of stay. Such prediction scores include Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, the Dense Artery, mRS, Age, Glucose, Onset-to-Treatment, and NIHSS (DRAGON) score, and the Totaled Health Risks in Vascular Events Stroke (THRIVE) score [24-26]. As summarised in previous reviews, with the exception of the Stroke Subtype, OCSP, Age, and Pre-stroke mRS (SOAR) score, all of these prognostic models have been developed to predict outcomes at or beyond 3 months post-stroke, rather than inpatient outcomes [27]. The SOAR score was developed to predict inpatient mortality and length of stay, but had limited performance (area under the receiver operator curve 0.79 for mortality and 0.61 for length of stay <8 days) [28]. A receiver operator curve is a commonly used metric in model performance analysis [29].

General Medicine is a field that specialises in undifferentiated and multisystem illnesses

General medicine, otherwise known as internal medicine, is a medical specialty that focusses on the care of individuals with undifferentiated illness or illnesses that affect multiple organ systems [30]. In addition to complex medical issues, general medicine may also provide care for individuals with complex social issues [31]. The role of a general medicine department may vary between centres. However, typically, general medicine comprises a large proportion of inpatient hospital admissions and the numbers of these admissions is increasing [32].

Discharge planning is the term used to describe the process by which multidisciplinary teams form plans around when a patient will leave hospital and which destination they will go to when they leave [33]. Effective discharge planning may reduce length of stay and readmissions, and accordingly has a vital role to play in improving bed flow and patient access to care [34]. This process is integral to general medicine due to both the medical and social complexity of the patients that are encountered and the potential for high patient numbers [35]. Part of the discharge planning process involves the generation of estimated discharge dates, which is the equivalent to an estimated length of stay, in advance of the time of discharge [36]. Estimated discharge date prediction is typically performed by a clinician and may be difficult to perform accurately.

The use of prognostic scores to aid with length of stay prediction, and other aspects of inpatient outcomes relevant to discharge planning (such as discharge destination), may be beneficial for general medicine. Scores developed for use on all hospital inpatients can be applied to the general medicine population. However, possibly due to the heterogeneity of this patient population, few scores specific to general medicine patients are currently available. Furthermore, the scores applied to all patients typically predict outpatient outcomes, or the only inpatient outcomes predicted are those of mortality or intensive care unit admission, such as in early warning scores [37]. Given that discharge planning most often involves patients being discharged to community dispositions (i.e., not the intensive care unit or death), and that the prediction of the timing of such discharges would assist with workflow planning, the development of such scores may be beneficial.

Machine learning may be able to be applied to assist with prognostication for discharge planning

Machine learning involves the use of computers to derive rules regarding data without human-defined feature selection [38]. Deep learning is a type of machine learning that involves the use of neural networks and their derivatives, such as convolutional neural networks [39]. Neural networks are a statistical model based on the structure of human neurons [40]. Machine learning, and in particular deep learning, has made significant advances in recent years with respect to the classification of medical information [41, 42]. It is likely that machine learning will have an increasing role in medicine, and may include a role in predicting inpatient outcomes relevant to discharge planning.

Bayes' theorem and Bayesian reasoning are important for both medical diagnosis and machine learning algorithms. Bayes' theorem provide a means of calculating conditional probabilities [43]. In medicine, this concept is important as it enables the pre-test probability of a diagnosis or outcome to be factored into the interpretation of a positive or negative test, with the subsequent determination of a post-test probability [44]. In machine learning, Bayes' theorem is also important in the probabilistic approach to modelling [45]. Accordingly, it can be seen this concept underpins the development and application of many medical machine learning models.

Machine learning may be applied to multiple data types, and even incorporate disparate data types into a single model [46]. For example, machine learning models can analyse discrete data fields (such as age, gender and smoking status), human speech and text, and images. When machine learning is applied to human text or speech this application is referred to as natural language processing [47].

One of the key components of analyses focussing on natural language processing is that of text mining. Text mining refers to the processing of text data so that it may be interpreted by subsequent statistical analyses, such as machine learning. Text mining in clinical medicine typically involves converting free-text notes and reports into structured data [48]. Text mining may employ a variety of strategies including named entity recognition, document clustering and word embeddings [49, 50].

After text data has been processed and is in a usable format, machine learning models may then be applied. There are many types of machine learning methodology. Different types of machine learning algorithms may be better suited to some tasks than others. Machine learning may encompass standard statistical methods, such as the use of logistic regression to predict binary outcomes on the basis of multiple explanatory variables [51]. However, when logistic regression is applied in a machine learning setting, the focus is typically on obtaining an accurate prediction of an outcome, as opposed to defining the relationships between variables [39]. Another method that may be used to predict binary outcomes is that of a decision tree. Clinicians are often familiar with the concept of a decision tree in the form of clinical guideline flowcharts. In the instance of machine learning, a series of hierarchical rules are defined, by which data are then classified [52]. Decision tree-based algorithms are often effective for unbalanced datasets. Random forest algorithms apply a combination of decision trees and have been shown to outperform traditional regression in certain circumstances [53]. Artificial neural networks are statistical models that involve, typically, multiple layers of interconnected nodes with weights that can be adjusted through backpropagation and may be particularly effective with large datasets [40]. All of these types of model are typically applied to supervised machine learning tasks, in which both the input and output labels are known and statistical associations can be made to predict the outcome in future cases in which the output is unknown [54].

The previously discussed models can be used to produce predictions, for which a number of performance metrics can be calculated. There is ongoing discussion in the medical and machine learning communities as to which performance metrics are optimal [55, 56]. Although the performance metrics for a given task may vary based upon the nature of that task, as a general rule it is likely that presenting a combination of performance metrics provides greater clarity than presenting only one metric. In classification tasks (e.g., “dead” vs “alive”, or “discharged” vs “not discharged”), performance metrics may be separated into two categories: prevalence-independent and prevalence-dependent. Prevalence-independent metrics include area under the receiver operator curve, sensitivity and specificity. Prevalence-dependent performance metrics include accuracy, positive predictive value and negative predictive value. In regression analyses, analyses in which the outcome to be predicted is a continuous (rather than categorical) variable, performance metrics may include mean absolute error, mean squared error, root mean squared error and R^2 values. There are some performance metrics that will be more familiar to clinicians, and others that are more familiar to machine learning researchers. A demonstration of this point is that the same metric may be referred to by different names depending on the field of the publication. For example, in some instances positive predictive value will be referred to as “precision”, and sensitivity will be referred to as “recall” [57]. There are also performance metrics that are used in machine learning literature that may be unfamiliar or infrequently used in medical literature. For example, an average precision may be calculated based upon a precision-recall curve. Another example of a metric commonly used in machine learning literature but not frequently in medical literature is that of the F1 score, which is the harmonic mean of precision and recall. Accordingly, the evaluation and presentation of a combination of these metrics is required in the development of machine learning models for clinical use.

The development of a machine learning model for clinical use can be thought of as similar to the steps involved in the development of a clinical decision rule or risk stratification scale. These steps have previously been described by Stiell et al. [58]. The steps include a pilot study (assessing feasibility), a derivation study (in which models are developed), validation studies (in which model performance is evaluated on external and prospective datasets), and implementation studies (in which the effect of model use are evaluated) [59].

However, even if a model has proved accurate in derivation and validation studies, if it does not predict outcomes and information that are clinically relevant, the model will be of limited value. For example, models that predict information that is already clear to a clinician or are extraneous to the diagnostic and therapeutic process may result in significant research resource expenditure without significant improvement to patient care. Accordingly, it is imperative that clinicians are involved in the development of medical machine learning models to ensure that they remain focussed on aiding with unaddressed questions that may improve patient care [60].

Outline and aims of thesis

The aim of this thesis is to develop and validate models for predicting clinically significant outcomes using machine learning methodologies, in particular deep learning, for patients admitted under Stroke and General Medicine units. Such clinically significant outcomes include timing of likely discharge and discharge destination. Key performance indicator data that was subsequently manually entered and clinical coding categorisation were also predicted. The objective is to develop such models so that they could result in healthcare efficiency savings and improve care delivery in future. The methods are presented in each chapter.

Specifically, this thesis aims to:

a) Provide a systematic review that highlights the current gaps in the literature with respect to length of stay prediction with machine learning in medical inpatients (**Chapter 2**)

b) Conduct a derivation study in which models are developed for the prediction of factors relevant to stroke discharge planning using discrete data fields (**Chapter 3**)

c) Conduct a prospective and external validation of the stroke inpatient models to predict factors relevant to discharge planning with discrete data fields **(Chapter 4)**

d) Investigate a variety of techniques, including natural language processing, to facilitate the automated collection of stroke key performance indicators from unstructured data (**Chapter 5**)

e) Investigate the application of natural language processing to stroke clinical coding (**Chapter 6**)

f) Investigate, in a general medicine population, the application of natural language processing to emergency department notes in order to predict length of stay (**Chapter 7**)

g) Conduct a derivation study in the general medicine population applying natural language processing to make recurrent daily predictions of discharge timing (**Chapter 8**)

h) Conduct the prospective and external validation of the general medicine recurrent length of stay prediction models (**Chapter 9**)

i) Provide an overall discussion, including future directions with respect to implementation studies (**Chapter 10**)

References

[1] Australian Institute of Health and Welfare. Hospital Activity. Australian Government; 2021.

[2] Birkmeyer JD, Barnato A, Birkmeyer N, Bessler R, Skinner J. The Impact Of The COVID-19 Pandemic On Hospital Admissions In The United States. Health Aff (Millwood). 2020;39:2010-7.

[3] Australian Medical Association. Public Hospitals: Cycle of Crisis. 2021.

[4] McPake B, Mahal A. Addressing the Needs of an Aging Population in the Health System: The Australian Case. Health Syst Reform. 2017;3:236-47.

[5] Australian Institute of Health and Welfare. Admitted Patients. Australian Government; 2021.

[6] Pefoyo AJ, Bronskill SE, Gruneir A, Calzavara A, Thavorn K, Petrosyan Y, et al. The increasing burden and complexity of multimorbidity. BMC Public Health. 2015;15:415.

[7] Campbell-Scherer D. Multimorbidity: a challenge for evidence-based medicine. Evidence-based medicine. 2010;15:165.

[8] Boyd CM, Kent DM. Evidence-based medicine and the hard problem of multimorbidity. J Gen Intern Med. 2014;29:552-3.

[9] Weiss CO, Varadhan R, Puhan MA, Vickers A, Bandeen-Roche K, Boyd CM, et al. Multimorbidity and evidence generation. J Gen Intern Med. 2014;29:653-60.

[10] J. W, Cook F, O'Day S, Peterson L, Wenger N, Reding D, et al. Relationship Between Cancer Patients' Predictions of Prognosis and Their Treatment Preferences. JAMA. 1998;279:1709-14.

[11] Deschepper M, Eeckloo K, Malfait S, Benoit D, Callens S, Vansteelandt S. Prediction of hospital bed capacity during the COVID- 19 pandemic. BMC Health Serv Res. 2021;21:468.

[12] Grood A, Blades K, Pendharkar S. A Review of Discharge-Prediction Processes in Acute Care Hospitals. Healthcare policy. 2016;12:105-15.

[13] Charlson M, Pompei P, Ales K, MacKenzie C. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of chronic diseases. 1987;40:373-83.

[14] Chassin M, Loeb J, Schmaltz S, Wachter R. Accountability Measures — Using Measurement to Promote Quality Improvement. The New England journal of medicine. 2010;363:683-8.

[15] Independent Hospital Pricing Authority. Australian Refined Diagnosis Related Groups (AR-DRGs). 2020.

[16] Feigin VL, Stark BA, Johnson CO, Roth GA, Bisignano C, Abady GG, et al. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology. 2021;20:795-820.

[17] Australian Bureau of Statistics. Causes of Death, Australia. Australian Government; 2020.

[18] Powers WJ. Acute Ischemic Stroke. N Engl J Med. 2020;383:252-60.

[19] Platz T. Evidence-Based Guidelines and Clinical Pathways in Stroke Rehabilitation-An International Perspective. Front Neurol. 2019;10:200.

[20] Hakkennes SJ, Brock K, Hill KD. Selection for inpatient rehabilitation after acute stroke: a systematic review of the literature. Arch Phys Med Rehabil. 2011;92:2057-70.

[21] Reeves MJ, Parker C, Fonarow GC, Smith EE, Schwamm LH. Development of stroke performance measures: definitions, methods, and current measures. Stroke. 2010;41:1573-8.

[22] Australian Commission on Safety and Quality in Health Care. Acute Stroke Clinical Care Standard. 2019.

[23] Urimubenshi G, Langhorne P, Cadilhac DA, Kagwiza JN, Wu O. Association between patient outcomes and key performance indicators of stroke care quality: A systematic review and meta-analysis. Eur Stroke J. 2017;2:287-307.

[24] Flint AC, Faigeles BS, Cullen SP, Kamel H, Rao VA, Gupta R, et al. THRIVE Score Predicts Ischemic Stroke Outcomes and Thrombolytic Hemorrhage Risk in VISTA. Stroke. 2013;44:3365-9.

[25] Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. Neurology. 2012;78:1916-22.

[26] Strbian D, Meretoja A, Ahlhelm F, Pitkaniemi J, Lyrer PA, Kaste M, et al. Predicting outcome of IV thrombolysis-treated ischemic stroke patients: The DRAGON score. Neurology. 2012;78:427-32.

[27] Drozdowska BA, Singh S, Quinn TJ. Thinking About the Future: A Review of Prognostic Scales Used in Acute Stroke. Front Neurol. 2019;10:274.

[28] Myint PK, Clark AB, Kwok CS, Davis J, Durairaj R, Dixit AK, et al. The SOAR (Stroke subtype, Oxford Community Stroke Project classification, Age, prestroke modified Rankin) score strongly predicts early outcomes in acute stroke. Int J Stroke. 2014;9:278-83.

[29] Hoo Z, Candlish J, Teare D. What is an ROC curve? Emerg Med J. 2017;34:357-9.

[30] Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. CMAJ Open. 2017;5:E842-E9.

[31] Nozzoli C, Mazzone A, Mathieu G, Iori I, Campanini M, La Regina M, et al. Complexity in hospital internal medicine departments: what are we talking about? Italian Journal of Medicine. 2013:142-55.

[32] Jenkins P, Thompson C, MacDonald A. What does the future hold for general medicine? Medical journal of Australia. 2011;195:49-50.

[33] Lin C-J, Cheng S-J, Shih S-C, Chu C-H, Tjung J-J. Discharge Planning. International Journal of Gerontology. 2012;6:237-40.

[34] Goncalves-Bradley DC, Lannin NA, Clemson LM, Cameron ID, Shepperd S. Discharge planning from hospital. Cochrane Database Syst Rev. 2016:CD000313.

[35] Ragavan MV, Svec D, Shieh L. Barriers to timely discharge from the general medicine service at an academic teaching hospital. Postgrad Med J. 2017;93:528-33.

[36] Ou L, Chen J, Young L, Santiano N, Baramy L, Hillman K. Effective discharge planning – timely assignment of an estimated date of discharge. Australian Health Review. 2011;35:357-63.

[37] Nannan Panday RS, Minderhoud TC, Alam N, Nanayakkara PWB. Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): A narrative review. Eur J Intern Med. 2017;45:20-31.

[38] Deo R. Machine Learning in Medicine. Circulation. 2015;123.

[39] Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19:64.

[40] Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS One. 2019;14:e0212356.

[41] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-8.

[42] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. Jama. 2016;316.

[43] Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380:1347-58.

[44] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24-9.

[45] Stiell IG, Wells GA. Methodologic Standards for the Development of Clinical Decision Rules in Emergency Medicine. Ann Emerg Med. 1999;33:437-47.

[46] Stiell IG, Bennett C. Implementation of Clinical Decision Rules in the Emergency Department. Academic Emergency Medicine. 2007;14:955-9.

[47] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195.

Chapter 2 - Machine Learning in the Prediction of Medical Inpatient Length of Stay, *Internal Medicine Journal*

Citation

Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T & Koblar S 2020, 'Machine Learning in the Prediction of Medical Inpatient Length of Stay', *Internal Medicine Journal*, https://doi.org/10.1111/imj.14962

Statement of Authorship

Title of Paper	Machine Learning in the Prediction of Medical Inpatient Length of Stay
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T & Koblar S 2020, 'Machine Learning in the Prediction of Medical Inpatient Length of Stay', *Internal Medicine Journal*, https://doi.org/10.1111/imj.14962

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Yiran Tan		
Contribution to the Paper	Data collection, critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Lauren Oakden-Rayner		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/3/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Background

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.

Aims

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.

Methods

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.

Results

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.

Conclusions

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.

Manuscript

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 (for example 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, 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.

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). By 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 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 CHARMS checklist [9] and 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.

Method

This review was constructed according to the PRIMSA-P 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 Appendix 1 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 length of stay 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 endpoint). LOS prediction for specific services during admission (e.g. LOS of time in 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] (see Appendix 2). 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 (see Figure 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, ten examined specific medical patient populations (see Table 1) 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] Eight studies included all inpatients at their respective centres, which encompassed medical patients [25-32] (see Table 2). Three studies included patients with acute kidney injury (AKI),[33] ICU admissions [34] and elective admissions,[35] 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 hours 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).

Many of the studies lacked a detailed description of study design elements according to the criteria in the employed critical appraisal framework. In particular, a number of 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 (for example, 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. Further, 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,[19] and one focussed on patients with unstable angina.[18] 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 Arabia.[16] This study employed a classification approach (< 3 days, 3-5 days and >5 days) and with a random forest model found an area under the receiver operator curve (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%, PPV 87.1%, NPV 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 hours), 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

Seven 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.[33-35] 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 hours inpatient admissions in ≥18 year-old patients (n = 216,221) to derive classification models predicting LOS ≥7 days or < 7 days.[31] 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.[31]

Three other studies also assessed models predicting all inpatient LOS as a classification task [27-29]. These studies reported accuracies of 78.5%,[27] 63.2% - 65.3% [28] and 97.3%.[29] Studies that evaluated LOS as a regression task reported mean absolute errors including 0.224,[26] 2.19 [30] and 4.68.[29] 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 (see 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 overfitting 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 hours, 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.

Conflicts of Interest

The authors declare they have no conflict of interest.

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.

[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. Academic Emergency Medicine. 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 Medicine. 2014;11.

[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. The Lancet. 2019;393:1577-9.

[12] Liu X, Faes L, Calvert MJ, Denniston AK. Extension of the CONSORT and SPIRIT statements. The Lancet. 2019.

[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.

[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 Systems with Applications. 2017;78:376-85.

[18] Huang Z, Dong W, Ji L, Duan H. Outcome Prediction in Clinical Treatment Processes. Journal of Medical Systems. 2015;40.

[19] 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.

[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 Applications2014. p. 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. European Journal of Internal Medicine. 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 Systems with Applications. 2013;40:6330-9.

[25] 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.

[26] 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.

[27] Livieris I, Kotsilieris T, Dimopoulos I, Pintelas P. Decision Support Software for Forecasting Patient's Length of Stay. Algorithms. 2018;11.

[28] 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 20182018. p. 16-22.

[29] 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.

[30] Cui L, Xie X, Shen Z, Lu R, Wang H. Prediction of the healthcare resource utilization using multi-output regression models. IISE Transactions on Healthcare Systems Engineering. 2019;8:291-302.

[31] 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.

[32] Liu V, Kipnis P, Gould M, Escobar G. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. Medical Care. 2010;48:739-44.

[33] 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.

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

[35] 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.

[36] Krstajic D, Buturovic L, Leahy D, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics. 2014;6:1-15.

Table 1**:** Studies predicting length of stay of medical inpatients from individual specialties

Citation	**Specialty**	**Retrospective vs prospective**	**Eligibility criteria**	**Sample size**	**Models used**	**LOS proportion or distribution**	**Regression or classification outcome**	**If LOS classification, what were the thresholds**	**Model performance**	**Critical Appraisal**
Tsai et al. 2016	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	Cardiology	Retrospective	All adult cardiology admissions.	16414	Random forest, artificial neural network, support vector machine and Baysian 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	Cardiology - CCF	Retrospective	All patients with admissions who had been diagnosed with CHF (although admission could be any cause)	20321	Regression tree (Cubist)	Mean LOS 6.24, Median 4, standard deviation 8.475.	Regression	NA	MAE 1. R2 0.79.	Few details on patient medical conditions. Clearly described approach to missing data.

Hachesu et al. 2013	Cardiology - IHD	Retrospective	All had coronary artery disease	2064	ANN, SVM and decision tree	LOS 0-5 days 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	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 2-3 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.
Mortona et al. 2014	Endocrinology	Retrospective	Not specified.	10000	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	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. NPV87.5	Clearly described LOS proportions in train/test sets. Clearly described train/test split.
Tsoukalas et al. 2015	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	Neurology - Stroke	Uncertain	Not specified.	716	Decision tree algorithm and Bayesian network	Uncertain	Classification	LOS 0-2 days, 3-7 days, 8-16	Accuracy 81.29%. AUC 0.936. Sensitivity 0.813. Specificity 0.896.	Uncertain proportion/ distribution of LOS. Uncertain inclusion criteria.

								days and >16 days		
Huang et al. 2013	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.

Table 2**:** Studies predicting length of stay 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, what were the thresholds**	**Model performance**	**Critical appraisal**
Steele & Thompson 2019	All elective admissions.	Retrospective	Not specified.	242024	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 & Ho 2019	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	All inpatient admissions.	Retrospective	Not specified.	100,000	disease+procedures2vec	Total dataset mean LOS 3.71 - 5.94	Regression	NA	R2 0.0766 - 0.4356	Diverse patient population. Limited discussion.
Caetano et al. 2015	All inpatient admissions.	Retrospective	Not specified.	26,431	Random forest	Uncertain	Regression	NA	R2 0.813, MAE 0.224, RMSE 0.469	Uncertain LOS distribution. Specified approach to missing data.
Livieris et al. 2018	All inpatient admissions.	Retrospective	Limited to > 65 year-olds	2,702	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. 2018b	All inpatient admissions.	Retrospective	Limited to > 65 year-olds	4,403	A variety of semi-supervised learning models were used including naieve 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	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 approach to missing data.
Cui et al. 2018	All inpatient admissions.	Retrospective	All admission except rare diagnoses.	750000	Multiple models including random forest, decision tree and neural network	Uncertain	Regression	NA	R2 0.554. RMSE 3.10. MAE 2.19.	Large sample size. Few details on patient medical conditions.
Liu et al. 2010	All inpatient admissions.	Retrospective	≥15 years old and not hospitalised for childbirth	155474	Logistic regression	Mean LOS 4.5 days +/- 7.7	Regression	NA	R2 0.146. MSE/1000 29.0.	Discussed exclusion of individuals with incomplete data. Included demographic details of patients.
Rajkomar et al. 2018	All inpatients.	Retrospective	≥18 years of age and ≥24 hour hospital admission	216221	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	Medicine - All patients in a trial with AKI.	Retrospective	Patients enrolled in AKI trial. Eligibility criteria as per AKI trial.	2,241	Random forest and logistic regression	Median LOS for whole cohort was 10.2 (6.0-17.2) days.	Regression	NA	R2 0.2 (0.14 - 026)	Included details on patient medical conditions. Discussed number of individuals screened for inclusion.

Appendix 1 – **Individual Database Search Strings**

PubMed:

("Machine learning"[All Fields] OR "artificial intelligence"[All Fields] OR "deep learning"[All Fields] OR "predictive analytics"[All Fields]) AND ("length of stay"[All Fields] OR "estimated discharge date"[All Fields] OR "length of hospital stay"[All Fields])

EMBASE:

('machine learning'/exp OR 'machine learning' OR 'artificial intelligence'/exp OR 'artificial intelligence' OR 'deep learning'/exp OR 'deep learning' OR 'predictive analytics') AND ('length of stay'/exp OR 'length of stay' OR 'estimated discharge date' OR 'length of hospital stay')

Web of Science:

ALL FIELDS:(("Machine learning" OR "artificial intelligence" OR "deep learning" OR "predictive analytics") AND ("length of stay" OR "estimated discharge date" OR "length of hospital stay"))

Timespan: All years. **Indexes:** SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC.

Appendix 2 – **Critical Appraisal Framework**

Prior to implementing the framework, address the following question: Is this a machine learning pilot/derivation/validation study aiming to predict a medical outcome? If yes, continue. If it is an implementation study, different critical appraisal methods will be required.

The following questions answered as either Yes/No/Unsure. *Significant additions as compared to the TRIPOD framework are italicised.*

- Title:
    - Does the title describe the area/results of the study?
- Abstract:
    - Does the abstract provide a summary of the aims, method and results of the study?
- Introduction:
    - Is the clinical relevance of the prediction of the target outcome explained?
    - Are the aims of the study stated?
    - Is a primary outcome (and secondary outcomes) defined? (e.g. AUC vs accuracy for desired outcome/output)
- Method:
    - Source of data:
        - Is the source of the data identified?

    - Are the dates from which the data was collected identified?
    - *Is it stated whether this data was prospectively/retrospectively collected?*
    - *If a validation study, is it stated that this data was collected separately from the data that was used to derive/develop the model initially?*
- Participants:
    - Are the eligibility criteria described?
- Outcome (Outputs):
    - Is it described which outcome is being predicted?
    - *Is it described how this outcome/output **label** was assessed/attributed?*
    - *Is the outcome/output **label** assessed/attributed in an objective and verifiable fashion?*
- Predictors (Inputs):
    - Is a complete list of the inputs used for the model provided?
    - If relevant, are there explanations of how these inputs were assessed?
- Sample size:
    - *Was the sample size of the training and test set adequate for the **purposes** of the study?*
- Missing data:
    - Is it explained how missing data was dealt with?
- Machine learning analysis:

    - *Is it stated which* ***programs*** *were used? (e.g. specify TensorFlow, Pytorch etc. ideally with version numbers)*
    - *Is the* ***preprocessing*** *of the data described?*
    - *If a* ***train/test split*** *was conducted, is it explicitly stated that this was performed only once?*
    - *Is the* ***type*** *and* ***architecture*** *of the models used in the study described? (in method or in results sections)*
    - *Is it described how training data was used to* ***develop and refine*** *the models? (e.g. K-fold cross-validation)*
    - *Is it described which data the model was applied to in order to produce the primary outcome? (i.e. what was the test set?)*
    - *If cut-off points were applied, is it described how these were selected?*
    - *Is it explicitly stated that performance analysis on the test set was performed only once?*
- Comparators
    - *If a comparator group was used, is the level of training of the members of the comparator group described?*
    - *If a comparator group was used, are the conditions under which the comparator group were performing described?*
    - *If a comparator score was used, is it described how the comparator score was calculated?*
- Statistical analysis:
    - *Are the statistical methods for the calculation of outcome metrics described? (e.g. method for calculating AUC, method*

        for calculating confidence intervals, method for calculating statistical significance tests)

  - Ethics:
    - *Is a statement regarding approval from local institutional review boards included? (with a reference number)*
- Results:
  - Participants
    - Are the demographic details of the patient population provided? (i.e. such that readers may determine whether results may be generalisable to their patient population)
    - *Are medical details of the patient population provided? (i.e. if all inpatients are included, provide information about the medical backgrounds and presenting complaints experienced by these patients)*
    - *Is it described how many patients were included in the study, as compared to how many were screened for inclusion?*
    - *Is adequate information presented, such that it is possible to determine the proportion/distribution of different classes/output results within the **total** dataset?*
  - Model training
    - *Is adequate information presented, such that it is possible to determine the proportion/distribution of different classes/output results within the **training** set?*
  - Model performance

        - *Is adequate information presented, such that it is possible to determine the proportion/distribution of different classes/output results within the* ***test*** *set?*
        - *Is a confusion matrix presented?*
        - *Are prevalence-****dependent*** *outcome metrics presented? (e.g. positive predictive value, negative predictive value, accuracy)*
        - *Are prevalence-****independent*** *outcome metrics presented? (e.g. AUC-ROC, sensitivity, specificity)*
        - *If cut-offs were employed, are these cut-off scores provided?*

- Discussion:
    - Is the discussion appropriate to the stage of development of the model? Does it discuss the model in the context of other studies and previous applications of this model?
    - Are the limitations of the study acknowledged?
    - Are reasonable suggestions for future research made?
- Conflicts of interest/funding:
    - Is there a statement regarding conflicts of interest/funding?

Chapter 3 - Stroke Prognostication for Discharge Planning with Machine Learning: A Derivation Study, *Journal of Clinical Neuroscience*

Citation

Bacchi S, Oakden-Rayner L, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Stroke Prognostication for Discharge Planning with Machine Learning: A Derivation Study', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2020.07.046

Statement of Authorship

Title of Paper	Stroke Prognostication for Discharge Planning with Machine Learning: A Derivation Study
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Oakden-Rayner L, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Stroke Prognostication for Discharge Planning with Machine Learning: A Derivation Study', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2020.07.046

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Lauren Oakden-Rayner		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/3/2022

Name of Co-Author	Prof David Menon		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	4/3/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Post-stroke discharge planning may be aided by accurate early prognostication. Machine learning may be able to assist with such prognostication. The study's primary aim was to evaluate the performance of machine learning models using admission data to predict the likely length of stay (LOS) for patients admitted with stroke. Secondary aims included the prediction of discharge modified Rankin Scale (mRS), in-hospital mortality, and discharge destination. In this study a retrospective dataset was used to develop and test a variety of machine learning models. The patients included in the study were all stroke admissions (both ischaemic stroke and intracerebral haemorrhage) at a single tertiary hospital between December 2016 to September 2019. The machine learning models developed and tested (75%/25% train/test split) included logistic regression, random forests, decision trees and artificial neural networks. The study included 2840 patients. In LOS prediction the highest area under the receiver operator curve (AUC) was achieved on the unseen test dataset by an artificial neural network at 0.67. Higher AUC were achieved using logistic regression models in the prediction of discharge functional independence (mRS ≤2) (AUC 0.90) and in the prediction of in-hospital mortality (AUC 0.90). Logistic regression was also the best performing model for predicting home vs non-home discharge destination (AUC 0.81). This study indicates that machine learning may aid in the prognostication of factors relevant to post-stroke discharge planning. Further prospective and external validation is required, as well as assessment of the impact of subsequent implementation.

Manuscript

Introduction

Accurate prognostication may aid post-stroke discharge planning [1]. However, accurately predicting likely outcomes for patients may be difficult, particularly early post-admission. Following admission, patient progress is typically noted over the course of several days, with physiotherapist and occupational therapist input, prior to forming discharge plans [2]. It is possible that machine learning may aid accurate prediction of post-stroke length of stay (LOS), disability, and discharge disposition based upon data available at admission. These predictions may aid in expediting discharge planning and facilitate improved patient care in a cost-effective manner.

Machine learning has previously been applied to small retrospective datasets to prognosticate stroke and assist discharge planning with variable success. The largest machine learning LOS prediction study used departmental data from 716 stroke inpatients and found an accuracy of 81.3% [3]. However, this study was limited by the inability to determine the distribution of LOS in their population. Several posters and abstracts have also presented results looking at stroke prognostication with machine learning in small patient groups (n = 66 and n = 158) with some promising results [4, 5]. Factors that have previously been shown to be associated with longer stroke LOS and/or greater disability include age, National Institute of Health Stroke Scale (NIHSS), heart failure and chronic kidney disease [6]. Another study has used machine learning to predict follow-up functional outcomes with an area under the receiver operator curve (AUC) of 0.888 for modified Rankin Scale prediction at 2-3 months following acute stroke [7].

The primary aim of this study was to evaluate the performance of machine learning models using time of admission data to predict (A) the likely LOS for patients admitted with stroke. Secondary aims included the prediction of (B) discharge modified Rankin Scale (mRS), (C) in-hospital mortality, and (D) discharge destination.

Materials and methods

Input and outcome data collection

Data were obtained from the existing stroke databases of a single tertiary hospital for all stroke admissions (both ischaemic stroke and intracerebral haemorrhage) between December 2016 to September 2019. Data were entered by stroke nurses or medical staff from patient notes and discharge summaries. Input data included: age, initial NIHSS, living alone status, estimated glomerular filtration rate, temperature, blood glucose level, blood pressure, pre-stroke mRS, sex, ethnicity, arrival method, ability to walk on arrival, result of swallow screening, receival of reperfusion therapy, socioeconomic status, and comorbidities including prior stroke, ischaemic heart disease, heart failure arrythmias, and active cancers. Postcode was utilised to estimate socioeconomic status using Australian Bureau of Statistics data [8].

Data pre-processing

Cases with missing data for any of the desired outcomes (Aims A-D) were excluded from analysis. Missing input data was replaced with median imputation. Prior to analysis, numerical data were standardised through feature scaling. Outcome data were dichotomised as follows: LOS >8 days vs ≤8 days (based on mean LOS of 8 days), mRS >2 or ≤2 (functional dependency or independence), discharge destination home vs non-home, and in-hospital mortality vs survival to hospital discharge. Data were randomly split into training and testing sets (75%/25%). This split was performed once.

Machine learning analysis

Machine learning analysis was conducted using open source Python libraries including TensorFlow (version 2.0) and SciKit-Learn (version 0.21.3).

Initially, univariate logistic regression was employed on the training set using all available input data to predict the dichotomised LOS outcome. The regression coefficients were ranked to estimate feature importance, and the six most important variables were selected for further experiments. Classification experiments were first employed on the training dataset using 5-fold cross-validation to develop and refine models. Logistic regression, random forest, decision tree and artificial neural network models were employed (see Supplementary Information 1 for additional information regarding the nature of these models). Similar methodology, with respect to the number of variables and machine learning models utilised, have been applied successfully in previous stroke prognostication machine learning studies and clinical decision rules [7, 9]. Using the results from this training data, model parameters were tuned, and architectures modified (see Supplementary Information 2). Following model development, the performance of each model was assessed on the previously unseen test dataset.

Statistical analysis

The AUC and average precision-recall scores were calculated with SciKit-Learn. As the models produce numeric predictions ranging from 0 to 1, we used Youden's index (determined during cross-validation on the training data) as a cut-off score for each model to produce the binary predictions [10]. High sensitivity and specificity (>90%) cut-off scores were calculated, and their performance evaluated, for the models with the highest AUC for each outcome.

Ethical Approval

The project received approval from the institutional Ethics Committee. Formal consent was not required for this type of study.

Results

Patient characteristics

We screened 2,922 patients for study inclusion, excluding those with missing outcome data (2,840 included). The mean age was 74.0 (SD 14.3 years), 1,259 were female (44.3%), median estimated pre-stroke mRS was 0 (IQR 0-1), and median admission NIHSS was 6 (IQR 3-15, range 0-42). The number of patients with ischaemic strokes was 2407 (84.8%) and 433 (15.2%) had haemorrhagic strokes. The median LOS was 4 (IQR 2 – 9, mean 8.1 days, SD 20.8 days, LOS ≤8 days in 2,124 individuals – 74.8%) and median discharge mRS was 3 (IQR 1-4, discharge mRS ≤2 in 1,381 individuals – 48.6%). In-hospital mortality was 15.9% (453 patients). A further 758 patients were discharged home (26.7%); remaining patients were discharged to non-home destinations, including rehabilitation and residential aged care facilities. Median imputation was required to replace 127 individual missing input datapoints.

Logistic regression feature selection

Initially logistic regression analysis was conducted using all available inputs (see *Method – Input and outcome data collection*). Factors most predictive of a length of stay >8 days were: ability to walk at time of admission, result of initial swallowing screening, pre-stroke mRS, age at time of admission, NIHSS at time of admission and socioeconomic status.

Classification experiments

The artificial neural network achieved the highest AUC in LOS dichotomised outcome prediction at 0.67 (see Table 1) (training set performance AUC 0.67 +/- SD 0.03). This performance was followed by the logistic regression model and random forest model with AUC of 0.66 and 0.64 respectively.

Significantly greater AUC were achieved in the prediction of discharge mRS. In this task, the highest AUC was achieved by the logistic regression model (AUC 0.90) (training set performance AUC 0.90 +/- SD 0.013) and artificial neural network (AUC 0.90). In the prediction of in-hospital mortality the best performing models were the logistic regression model (AUC 0.90) (training set performance AUC 0.90 +/- SD 0.022) and the artificial neural network (AUC 0.85). Logistic regression and artificial neural networks were more accurate in predicting home vs non-home discharge destination (both AUC 0.81) (logistic regression training set performance AUC 0.79

+/- SD 0.013) than random forest and decision tree models, which achieved lower AUC (accuracy 0.77 and 0.64 respectively).

When high sensitivity and specificity cut-offs were applied, different levels of performance were demonstrated (see Table 2). Of note, the prediction of discharge mRS was able to maintain a sensitivity of 0.72, while providing specificity of 0.90.

Discussion

This study has shown that, using data available at the time of admission, machine learning may aid in predicting aspects of stroke patient prognosis relevant to discharge planning. In particular, logistic regression and artificial neural networks were able to successfully predict discharge mRS, in-hospital mortality and discharge to home. However, accurate prediction of LOS proved challenging.

The models in this study have comparable performance to other stroke prognostication scores such as ASTRAL, DRAGON, FSV, iSCORE, PLAN and THRIVE. These scales, which have been reviewed previously [11], have been shown to have a range of AUC from 0.71 to 0.89 for predicting a variety of functional and mortality outcomes at 3-12 months post-stroke. The only scale derived primarily for predicting similar inpatient outcomes to those in this study, the SOAR and modified-SOAR scores, has been shown to have 0.79 AUC for predicting inpatient mortality and 0.61 AUC for predicting LOS < 8 days [12].

The most likely explanation as to why LOS was unable to be more effectively predicted in this study is that factors significantly affecting LOS were not represented in the available input data. Such factors may include the insurance status of the patients (public vs private), recurrent inpatient stroke or stroke progression (e.g. haematoma expansion), weekend/out-of-hours admissions [13], and the bed-state at the time of admission (i.e. system overcrowding and access block). Although admission NIHSS accurately measures stroke severity, imaging data (ASPECTS and ischaemic core calculation) and successful reperfusion may enhance outcome prediction [14]. Further, in addition to baseline data, data collected *throughout* the admission, with landmark analysis [15], may refine the original prediction [16].

A limitation of this study is that it was conducted at a single centre. It should be noted that in the test set for the in-hospital mortality classification task 83% of patients survived. Still, it can be seen that the logistic regression model was not simply predicting the most common class, since it achieved sensitivity of 0.81, specificity of 0.86 and average precision-recall score of 0.98. The inclusion of additional information such as 'out-of-hours' admission, baseline neuroimaging, insurance status and a measure of healthcare system bed-state may have improved model performance.

Our study supports the need for further research in this area. Future studies should aim to improve existing models, and prospectively validate such models. Implementation studies are also required to demonstrate benefits in patient or healthcare-system oriented outcomes with model deployment.

Conflict of Interest

The authors declare that there is no conflict of interest.

Sources of support

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Cho JS, Hu Z, Fell N, Heath GW, Qayyum R, Sartipi M. Hospital Discharge Disposition of Stroke Patients in Tennessee. South Med J. 2017;110:594-600.

[2] Luker J, Bernhardt J, Grimmer K, Edwards I. A qualitative exploration of discharge destination as an outcome or a driver of acute stroke care. BMC health services research. 2014;14:193.

[3] 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.

[4] Arndt S, Bennett G, Wojcik K, Albar A, Alhasan M, Ma J, et al. Length of stay in mechanical thrombectomy, and machine learning improvement of predictive analysis. J NeuroIntervent Surg. 2017;9:A1–A94.

[5] Globas C, Zenko B, Luft AR. Predicting disability and quality of life after ischemic stroke: A machinelearning-based analysis from the zurich observational registry of rehabilitation outcomes (“ZORRO”). Cerebrovascular Diseases. 2012;33 SUPPLS. 2:529-30.

[6] Mohamed W, Bhattacharya P, Shankar L, Chaturvedi S, Madhavan R. Which Comorbidities and Complications Predict Ischemic Stroke Recovery and Length of Stay? Neurologist. 2015;20:27-32.

[7] Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. Stroke. 2019;50:1263-5.

[8] Statistics ABo. Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016. Commonwealth Goverment of Australia; 2016.

[9] Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. Neurology. 2012;78:1916-22.

[10] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32-5.

[11] Drozdowska BA, Singh S, Quinn TJ. Thinking About the Future: A Review of Prognostic Scales Used in Acute Stroke. Front Neurol. 2019;10:274.

[12] Myint PK, Clark AB, Kwok CS, Davis J, Durairaj R, Dixit AK, et al. The SOAR (Stroke subtype, Oxford Community Stroke Project classification, Age, prestroke modified Rankin) score strongly predicts early outcomes in acute stroke. Int J Stroke. 2014;9:278-83.

[13] Turner M, Barber M, Dodds H, Dennis M, Langhorne P, Macleod MJ, et al. Stroke patients admitted within normal working hours are more likely to achieve process standards and to have better outcomes. J Neurol Neurosurg Psychiatry. 2016;87:138-43.

[14] Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes: A Pilot Study. Academic Radiology. 2019.

[15] Morgan CJ. Landmark analysis: A primer. J Nucl Cardiol. 2019;26:391-3.

[16] Huang Z, Juarez JM, Duan H, Li H. Length of stay prediction for clinical treatment process using temporal similarity. Expert Systems with Applications. 2013;40:6330-9.

Table 1**:** Results of classification experiments predicting stroke outcomes with data available at the time of admission

Outcome	**Model**	**AUC**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F1 Score**	**PR average score**	**Accuracy**
LOS	Artificial neural network	**0.67**	305	219	132	54	0.58	0.71	0.85	0.38	0.69	0.84	0.62
	Decision tree	0.56	394	130	69	117	0.75	0.37	0.77	0.35	0.76	0.76	0.65
	Logistic regression	0.66	336	188	116	70	0.64	0.62	0.83	0.38	0.72	0.83	0.64
	Random Forest	0.64	225	299	148	38	0.43	0.80	0.86	0.33	0.57	0.84	0.53
Discharge mRS	Artificial neural network	0.9	291	55	305	59	0.84	0.84	0.83	0.85	0.84	0.88	0.84
	Decision tree	0.75	250	96	281	83	0.72	0.77	0.75	0.75	0.74	0.68	0.75
	Logistic regression	**0.90**	291	55	305	59	0.84	0.84	0.83	0.85	0.84	0.88	0.84
	Random Forest	0.88	299	47	284	80	0.86	0.78	0.79	0.86	0.82	0.87	0.82
In-hospital mortality	Artificial neural network	0.85	509	83	78	40	0.86	0.66	0.93	0.48	0.89	0.97	0.83
	Decision tree	0.66	540	52	48	70	0.91	0.41	0.89	0.48	0.90	0.88	0.83
	Logistic regression	**0.90**	477	115	102	16	0.81	0.86	0.97	0.47	0.88	0.98	0.82

	Random Forest	0.88	480	112	97	21	0.81	0.82	0.96	0.46	0.88	0.97	0.81
Discharge destination	Artificial neural network	0.81	148	44	382	136	0.77	0.74	0.52	0.90	0.62	0.56	0.75
	Decision tree	0.64	96	96	406	112	0.50	0.78	0.46	0.81	0.48	0.37	0.71
	Logistic regression	**0.81**	146	46	393	125	0.76	0.76	0.54	0.90	0.63	0.57	0.76
	Random Forest	0.77	169	23	290	228	0.88	0.56	0.43	0.93	0.57	0.53	0.65

AUC = area under the receiver operator curve, TP = True positives, FN = false negatives, TN = true negatives, FP = false positives, PPV = positive predictive value, NPV = negative predictive value, PR = Precision-Recall, mRS = modified Rankin Scale, LOS = Length of stay

Table 2**:** Results of classification experiments applying high sensitivity/specificity cut-off scores to models developed to predict predict stroke outcomes

Outcome	**Model and cut-off**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
LOS	Artificial neural network – high specificity	0.27	0.90	0.89	0.30
	Artificial neural network – high sensitivity	0.91	0.20	0.76	0.44
Discharge mRS	Logistic regression – high specificity	0.72	0.90	0.87	0.77
	Logistic regression – high sensitivity	0.90	0.72	0.75	0.88
In-hospital mortality	Logistic regression – high specificity	0.72	0.91	0.97	0.39
	Logistic regression – high sensitivity	0.90	0.65	0.93	0.57
Discharge destination	Logistic regression – high specificity	0.42	0.90	0.61	0.81
	Logistic regression – high sensitivity	0.90	0.49	0.39	0.93

PPV = positive predictive value, NPV = negative predictive value, mRS = modified Rankin Scale, LOS = Length of stay

Supplementary Information 1 - **Background information regarding machine learning models**

Machine learning encompasses a wide variety of techniques, which can be categorised broadly into three groups: supervised, semi-supervised and un-supervised. Supervised machine learning involves datasets where the “ground truth” is known for all input and output datapoints. For example, this study could be considered an example of supervised machine learning because the desired outcome (e.g. discharge destination or length of stay) was known for all individuals included in the study.

When conducting supervised machine learning studies, there are a wide variety of models that may be used to predict the desired outcome. One way to categorise these models is as deep learning vs non-deep learning. Non-deep learning models include K-nearest neighbours, decision tree, Naïve Bayes and support vector machine algorithms. Deep learning refers to the use of machine learning models based upon artificial neural networks. The four types of models employed in the project were logistic regression, decision tree, random forest and artificial neural networks.

Logistic regression is a statistical model that uses the logistic (sigmoid) function to predict the probability of an observation being associated with a given class. There are a variety of types of logistic regression, including binary, multinomial and ordinal logistic regression.

Decision tree algorithms, when referring to classification trees, employ a method involving sequential divisions of data aiming to minimise entropy, that eventually yields the division of a dataset into separate classes. These sequential divisions of data can be visualised as a “tree” or flowchart of decisions that outline the separate steps that were taken while dividing the data into separate classes.

A random forest model employs multiple decision trees (see above), which are each developed on a random subset of data, to classify observations.

Artificial neural networks use a sequence of layers of nodes that are interconnected with different weights, to predict the classification of a given observation. When training an artificial neural network, predictions are made, based upon the current values of the weights that join the nodes, and then a cost function is used to calculate an indicator of the amount of error in the predictions. This value is then used to update/refine the weights of the network (backpropagation) with the aim of minimising the cost function. The architecture of a network refers to the number of layers and numbers of nodes in the network, as well as other factors that may be modified such as the activation function of different nodes, and inclusion of layers with different functions (such as dropout layers).

Supplementary Information 2 - **ANN architecture and hyperparameter tuning**

The architecture of the ANN that was used in the classification experiments had only 2 fully connected layers. Following the input layer, there was a dense layer with 10 nodes, then a second dense layer with 4 nodes. After this followed the output layer with 1 node. The loss function that was employed was binary cross-entropy. No dropout layers were used.

Hyperparameters were tuned using a grid-search function. Hyperparameters that were tuned were batch size, number of epochs and learning rate. Ultimately the ANN that was used for the classification experiments on the test set used a batch size of 300, 100 epochs and a learning rate of 0.01.

Chapter 4 - Prospective and external validation of stroke discharge planning machine learning models, *Journal of Clinical Neuroscience*

Citation

Bacchi S, Oakden-Rayner L, Menon D, Moey A, Jannes J, Kleinig T & Koblar S 2021, ‘Prospective and external validation of stroke discharge planning machine learning models’, *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.12.031

Statement of Authorship

Title of Paper	Prospective and external validation of stroke discharge planning machine learning models
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Oakden-Rayner L, Menon D, Moey A, Jannes J, Kleinig T & Koblar S 2021, 'Prospective and external validation of stroke discharge planning machine learning models', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.12.031

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Lauren Oakden-Rayner		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/3/2022

Name of Co-Author	Prof David Menon		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	4/3/2022

Name of Co-Author	Dr Andrew Moey		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	4/3/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Machine learning may be able to help with predicting factors that aid in discharge planning for stroke patients. This study aims to validate previously derived models, on external and prospective datasets, for the prediction of discharge modified Rankin scale (mRS), discharge destination, survival to discharge and length of stay. Data were collected from consecutive patients admitted with ischaemic or haemorrhagic stroke at the Royal Adelaide Hospital from September 2019 to January 2020, and at the Lyell McEwin Hospital from January 2017 to January 2020. The previously derived models were then applied to these datasets with three pre-defined cut-off scores (high-sensitivity, Youden's index, and high-specificity) to return indicators of performance including area under the receiver operator curve (AUC), sensitivity and specificity. The number of individuals included in the prospective and external datasets were 334 and 824 respectively. The models performed well on both the prospective and external datasets in the prediction of discharge mRS ≤2 (AUC 0.85 and 0.87), discharge destination to home (AUC 0.76 and 0.78) and survival to discharge (AUC 0.91 and 0.92). Accurate prediction of length of stay with only admission data remains difficult (AUC 0.62 and 0.66). This study demonstrates successful prospective and external validation of machine learning models using six variables to predict information relevant to discharge planning for stroke patients. Further research is required to demonstrate patient or system benefits following implementation of these models.

Manuscript

Introduction

Discharge planning is an important part of inpatient stroke case evaluation, facilitating efficient and effective care, as well as accurate bed-state planning [1]. Machine learning may be able to help with predicting factors that aid with this task. In a previous study, machine learning models were derived that successfully predicted factors relevant to discharge planning [2]. However, prior to the implementation of any such algorithms, validation studies are required to confirm the performance observed in the derivation study.

The development of machine learning algorithms for use in clinical medicine may be thought of as similar to the stages of development of clinical decision rules. These stages include a pilot study (to assess feasibility), a derivation study (to develop the model), external/prospective validation studies (to validate the performance of the model seen in the derivation study on data separate to the derivation dataset) and implementation studies (to demonstrate improvement in patient or system outcomes from use of the model) [3]. Currently there are many clinical machine learning pilot and derivation studies, but relatively few authors have conducted prospective/external validation or implementation studies of their models.

The aim of this study was to evaluate the performance of previously derived machine learning models, for the prediction of stroke outcomes relevant to discharge planning, in a prospectively collected dataset and a separate dataset collected from an external centre. The outcomes for which this evaluation was conducted were discharge modified Rankin scale (mRS), discharge destination, survival to discharge and length of stay (LOS). The primary outcome was the area under the receiver operator cure (AUC) for predicting discharge mRS.

Materials and methods

Participating centres

The two South Australian hospitals involved in the project were the Royal Adelaide Hospital (Location 1) and the Lyell McEwin Hospital (Location 2). The Royal Adelaide Hospital is the sole Comprehensive Stroke Centre in Adelaide. The Lyell McEwin hospital is a northern Adelaide tertiary hospital, which is a Primary Stroke Centre without on-site neurosurgery or endovascular thrombectomy.

Machine learning models

The models were originally derived using data from the Royal Adelaide Hospital from between December 2016 to September 2019 (not including September) as described previously [2]. The models include logistic regression models for predicting discharge mRS, discharge destination, and survival to discharge; and an artificial neural network for prediction of LOS. There are six inputs required for the models, namely: age at time of admission, National Institutes of Health Stroke Scale (NIHSS) at time of admission, ability to walk at time of admission, result of initial swallowing screening, pre-stroke mRS and socioeconomic status (as estimated from postcode using Australian Bureau of Statistics data) [4]. In the prediction of each outcome, the individual variables were weighted differently. For example (noting the application of feature scaling), in the prediction of discharge destination, regression coefficients for the variables were age -0.22, NIHSS -0.58, socioeconomic status -0.02, pre-stroke mRS ranged from 0.13 to -0.15, inability to walk at time of admission -0.28 and failed swallow screen -0.21.

Data collection

Two sets of data were collected for the validation of the machine learning models. Data from both sites included information on consecutive stroke admissions extracted by stroke nursing, medical and allied health staff from medical records using a standardised form. Data were collected for the prospective validation of the models at the Royal Adelaide Hospital (Location 1) between September 2019 to January 2020. Data for the external validation of the models were collected from the Lyell McEwin Hospital (Location 2) from between January 2017 to January 2020.

Data pre-processing

Individuals for which the outcome data was missing were excluded from analysis. The outcomes for prediction were discharge functional dependence (mRS >2) vs independence (≤2), home vs non-home discharge destination, survival to hospital discharge and LOS >8 days vs ≤8 days. Median imputation was used to replace missing input data. Feature scaling was applied.

Statistical and performance analysis

Validation cohort demographics were compared using unpaired t-tests and chi-squared tests. The previously derived models were applied to the two datasets to generate performance metrics. Performance metrics included prevalence-independent (AUC, sensitivity and specificity) and prevalence-dependent metrics (positive predictive value, negative predictive value and accuracy), calculated as standard and using SciKit-Learn (version 0.21.3). For each model, performance was evaluated with three previously defined cut-off scores (high-sensitivity, Youden's index from the derivation study, and high-specificity).

Performance of the model for validation was also specifically evaluated in a number of subgroups including individuals with intracerebral haemorrhage, large vessel occlusion, and those receiving thrombolysis and endovascular thrombectomy. To help evaluate the performance of the model for validation on these subgroups, new models were derived on the previous derivation dataset using data only from these subgroups, as described previously [2]. The performance of these subgroup models for prediction of the primary outcome was evaluated in the prospective and external validation datasets.

Ethical Approval

The project received approval from institutional Ethics Committees, with waiver of individual consent.

Results

Patient characteristics

The number of individuals included at Location 1 was 334 (98.8 % of the total 338 screened for inclusion) and, at Location 2, 824 (77.4% of the total 1065 screened for inclusion). At Location 1 there was a mean age 75.2 years (SD 13.1), 151 female individuals (45.2%), 82 discharged to home (24.6%), 285 survived to discharge (85.3%), mean LOS 6.4 days (SD 8.1), 163 individuals with mRS at discharge ≤2 (48.8%), admission NIHSS median 5 (IQR 3-12), and 284 ischaemic strokes (85.0%).

At Location 2 there was a mean age of 73.3 years (SD 13.6), 371 females (45.0%), 294 discharged to home (35.7%), 747 survived to discharge (90.7%), LOS 7.4 days (SD 12.4), 354 individuals with mRS at discharge ≤2 (43.0%), admission NIHSS median 4 (IQR 2-8) and 709 ischaemic strokes (86.0%). There were several statistically significant differences between the included individuals from Location 1 and Location 2, including a higher proportion of acute strokes (symptom noted to arrival time <24 hours 86.5% vs 68.1%, p <0.01) and a higher proportion with moderate-severe strokes (NIHSS ≥6 47.9% vs 36.4%, p <0.01). For additional details regarding characteristics of individuals, please see Table 1.

Prospective validation – Location 1

In the primary outcome of AUC for prediction of discharge mRS ≤ 2 a performance of 0.85 was achieved (see Table 2). Using the high-sensitivity cut-off, a sensitivity of 0.91 was able to be attained while maintaining a specificity of 0.64. Conversely, with the high-specificity cut-off, a specificity of 0.82 was achieved with a sensitivity of 0.69. When examining instances of misclassification following the application of the Youden's index cut-off (74), it was found that 14.9% (11/74) of misclassified cases had intracerebral haemorrhage, 85.1% (63/74) had ischaemic stroke, 40.5% (30/74) had non-large vessel occlusion ischaemic stroke, 44.6% (33/74) had large vessel occlusion ischaemic stroke, 18.9% (14/74) received thrombolysis and 14.9% (11/74) received endovascular thrombectomy. In the prediction of home discharge destination and survival to discharge the models returned AUC of 0.76 and 0.91 respectively. The prediction of LOS was less accurate, with an AUC of 0.62.

When performance on subgroups was analysed, performance in the prediction of discharge mRS varied from AUC 0.76 (intracerebral haemorrhage) to 0.86 (ischaemic strokes with large vessel occlusion) (see Table 3). The intracerebral haemorrhage subgroup for the Location 1 cohort was relatively small, and the proportion of individuals that had a mRS ≤2 at the time of discharge was 15.7% (8/51). Subgroup models derived on data from patients with intracerebral haemorrhage (AUC 0.83), large vessel occlusion (AUC 0.85), thrombolysis (AUC 0.74), and endovascular thrombectomy (AUC 0.86) had reasonable performance when applied to the Location 1 dataset.

External validation – Location 2

The performance on the external dataset was similar to that of the model performance on the prospective dataset. In the prediction of discharge mRS ≤ 2 the model achieved an AUC of 0.87 (see Table 2). Using high-sensitivity and high-specificity cut-offs, sensitivity of 0.89 (with specificity 0.68) and specificity of 0.88 (with sensitivity 0.60) were achieved. After the application of the Youden's index cut-off, it was found that in the 177 cases that were misclassified by this model, 10.7% (19/177) had intracerebral haemorrhage, 89.3% (158/177) had ischaemic stroke, 68.9% (122/177) had non-large vessel occlusion ischaemic stroke, 20.3% (36/177) had large vessel occlusion ischaemic stroke, 13.6% (24/177) received thrombolysis and 4.5% (8/177) received endovascular thrombectomy. In this Location 2 cohort the percentage of individuals with intracerebral haemorrhage with a mRS ≤2 at the time of discharge was 31.9% (38/119). The AUC achieved for the prediction of discharge destination, survival to discharge and LOS were all similar or slightly higher on the external validation dataset (AUC 0.78, 0.92 and 0.66 respectively) compared to the prospective validation dataset.

In the analysis of performance in the prediction of mRS on subgroups, AUC varied from 0.85 (non-large vessel occlusion ischaemic strokes) through to 0.91 (large vessel occlusion ischaemic strokes). When the subgroup models were applied to the Location 2 dataset an AUC of 0.90 was returned for patients with intracerebral haemorrhage, 0.91 for large vessel occlusion, 0.84 for individuals who received thrombolysis and 0.89 for those who received endovascular thrombectomy.

Discussion

This study has demonstrated that the previously derived models [2] for the prediction of stroke outcomes to aid in discharge planning had similar performance on prospective and external validation datasets. These results included sound performance in the prediction of discharge mRS, discharge destination and survival to discharge. However, predicting LOS remains difficult.

The consistent difficulty in predicting LOS with only admission data is felt likely to reflect the heterogeneity of a patient’s course following admission, for example, variation in reperfusion outcomes in ischaemic stroke. In areas outside of stroke, more accurate estimates of LOS have been achieved using recurrent predictions that take into account further information regarding a patient’s course collected on an ongoing basis throughout their admission [5, 6]. However, for such a model to be feasible with respect to possible future implementation, integration into an electronic health record would be required to automate the data collection required for such recurrent predictions. Means of streamlining integration of machine learning models into electronic health records is likely to be a key area for future research and development in its own right, in addition to facilitating the development of models themselves and the deployment of such models [7, 8]. The utilisation of alternative cut-off values (for example, greater or less than a two-day stay) is another strategy that could be employed to improve the accuracy of LOS predictions. The use of alternative cut-off values could also be incorporated with the generation of recurrent predictions.

The models in this study are focussed on discharge planning. Discharges are only one aspect of patient movement within a hospital that contribute to overall demand and patient flow [9]. Other aspects of patient movement that may be able to be predicted with machine learning include movement from the emergency department and transfer to the intensive care unit [9, 10].

Although the models have shown good external validity in this study, it is acknowledged that aspects of discharge planning may be centre-dependent. Such facility-specific factors may include bed availability at rehabilitation centres, rehabilitation in the home capacity, weekend staffing and insurance frameworks. Even so, the use of a given model across different centres may facilitate comparative audits.

There may be diverse centres in significantly different healthcare systems where the models perform less effectively. If this were demonstrated to be the case, it may be hypothesised that the use of local data for the six variables used in this study may enable the development of centre-specific algorithms. In comparison to more complex models, for which all required input data may not be available, the models in this study make this task, and comparative audits, feasible. Even with relatively simple models, a significant number of individuals may require exclusion due to incomplete outcome data, as was the case for Location 2, which is a limitation of this study. It should be noted that when Australian Bureau of Statistics socioeconomic data is updated in future this change would require the step that derives this value from postcode to be performed with the updated data [4]. However, the model itself would not require change. In addition to developing centre-specific models with local data, other strategies to improve model performance could include using additional input data (such as system related information with respect to bed-state or computed tomography perfusion and angiography imaging), using input data collected throughout the admission (as opposed to only at the time of admission) and developing models that target specific subgroups of stroke patients.

This study has demonstrated external and prospective validation of the previously derived models. Further external validation at interstate or international centres may be beneficial. Before use in clinical practice, a further study demonstrating improvement in patient or system outcomes following implementation of the models is required.

Conflict of Interest

The authors declare that there is no conflict of interest.

Sources of support

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Andrew NE, Busingye D, Lannin NA, Kilkenny MF, Cadilhac DA. The Quality of Discharge Care Planning in Acute Stroke Care: Influencing Factors and Association with Postdischarge Outcomes. J Stroke Cerebrovasc Dis. 2018;27:583-90.

[2] Bacchi S, Oakden-Rayner L, Menon DK, Jannes J, Kleinig T, Koblar S. Stroke prognostication for discharge planning with machine learning: A derivation study. Journal of Clinical Neuroscience. 2020;79:100-3.

[3] Stiell IG, Wells GA. Methodologic Standards for the Development of Clinical Decision Rules in Emergency Medicine. Ann Emerg Med. 1999;33:437-47.

[4] Australian Bureau of Statistics. Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016. Commonwealth Goverment of Australia; 2016.

[5] Huang Z, Juarez JM, Duan H, Li H. Length of stay prediction for clinical treatment process using temporal similarity. Expert Systems with Applications. 2013;40:6330-9.

[6] Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, et al. Mixed-data deep learning in repeated predictions of general medicine length of stay: a derivation study. Intern Emerg Med. 2021;16:1613-17.

[7] Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25:1419-28.

[8] Davenport T, Hongsermeier T, McCord K. Using AI to Improve Electronic Health Records. Harvard Business Review; 2018.

[9] El-Bouri R, Taylor T, Youssef A, Zhu T, Clifton DA. Machine learning in patient flow: a review. Progress in Biomedical Engineering. 2021.

[10] El-Bouri R, Eyre DW, Watkinson P, Zhu T, Clifton DA. Hospital Admission Location Prediction via Deep Interpretable Networks for the Year-Round Improvement of Emergency Patient Care. IEEE J Biomed Health Inform. 2021;25:289-300.

Table 1**:** Comparison of cohort demographics in prospective and external validation datasets

Characteristic	**Location 1 (prospective validation) (n = 334)**	**Location 2 (external validation) (n = 824)**	**Statistical significance (p value)**
Mean age (standard deviation)	75.2 (13.1)	73.3 (13.6)	0.029
Number female (%)	150 (44.9%)	369 (44.8%)	0.98
Number with moderate-severe strokes (NIHSS ≥6) (%)	160 (47.9%)	300 (36.4%)	<0.01
Number acute strokes (symptom-noted to door time <24 hours) (%)	289 (86.5%)	561 (68.1%)	<0.01
Number with ischaemic stroke (%) (vs opposed to intracerebral haemorrhage)	283 (84.7%)	705 (85.6%)	0.79
Number with large vessel occlusion (%)	149 (44.6%)	185 (22.5%)	<0.01
Number received thrombolysis (%)	38 (11.4%)	70 (8.5%)	0.16
Number received endovascular thrombectomy (%)	54 (16.2%)	30 (3.6%)	<0.01
Number with hypertension (%)	221 (66.2%)	594 (72.1%)	0.05

Number with diabetes mellitus (%)	69 (20.7%)	247 (30.0%)	<0.01
Number with ischaemic heart disease (%)	51 (15.3%)	187 (22.7%)	<0.01
Number with active smoking (%)	42 (12.6%)	142 (17.2%)	0.06

NIHSS = National Institutes of Health Stroke Scale.

Table 2**:** Results of prospective and external validation of previously derived stroke discharge planning machine learning models

Outcome	**AUC**	**PR score**	**Cut-off employed**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F1 Score**	**Accuracy**
Location 1 (prospective validation)													
Discharge mRS ≤2	0.85	0.83	High-specificity	113	50	140	31	0.69	0.82	0.78	0.74	0.74	0.76
			Youden's index	130	33	130	41	0.80	0.76	0.76	0.80	0.78	0.78
			High-sensitivity	148	15	110	61	0.91	0.64	0.71	0.88	0.80	0.77
Discharge destination was to home	0.76	0.46	High-specificity	18	64	230	22	0.22	0.91	0.45	0.78	0.30	0.74
			Youden's index	57	25	177	75	0.70	0.70	0.43	0.88	0.53	0.70
			High-sensitivity	76	6	109	143	0.93	0.43	0.35	0.95	0.50	0.55
Survival to discharge	0.91	0.98	High-specificity	215	69	45	5	0.76	0.90	0.98	0.39	0.85	0.78
			Youden's index	239	45	42	8	0.84	0.84	0.97	0.48	0.90	0.84
			High-sensitivity	254	30	36	14	0.89	0.72	0.95	0.55	0.92	0.87

LOS ≤8 days	0.62	0.85	High-specificity	51	206	67	10	0.20	0.87	0.84	0.25	0.32	0.35
			Youden's index	144	113	51	26	0.56	0.66	0.85	0.31	0.67	0.58
			High-sensitivity	222	35	12	65	0.86	0.16	0.77	0.26	0.82	0.70
Location 2 (external validation)													
Discharge mRS ≤2	0.87	0.88	High-specificity	280	190	310	44	0.60	0.88	0.86	0.62	0.71	0.72
			Youden's index	358	112	290	64	0.76	0.82	0.85	0.72	0.80	0.79
			High-sensitivity	420	50	239	115	0.89	0.68	0.79	0.83	0.84	0.80
Discharge destination was to home	0.78	0.63	High-specificity	56	238	503	27	0.19	0.95	0.67	0.68	0.30	0.68
			Youden's index	177	117	412	118	0.60	0.78	0.60	0.78	0.60	0.71
			High-sensitivity	270	24	239	291	0.92	0.45	0.48	0.91	0.63	0.62
Survival to discharge	0.92	0.99	High-specificity	518	229	73	4	0.69	0.95	0.99	0.24	0.82	0.72
			Youden's index	607	140	68	9	0.81	0.88	0.99	0.33	0.89	0.82
			High-sensitivity	667	80	60	17	0.89	0.78	0.98	0.43	0.93	0.88

LOS ≤8 days	0.66	0.81	High-specificity	108	483	212	21	0.18	0.91	0.84	0.31	0.30	0.39
			Youden's index	343	248	157	76	0.58	0.67	0.82	0.39	0.68	0.61
			High-sensitivity	530	61	49	184	0.90	0.21	0.74	0.45	0.81	0.70

Table 3**:** Sub-group analyses demonstrating model performance in the prediction of dichotomised discharge modified Rankin scale (≤2) using Youden's index cut-off score

Subgroup	**Location 1 (prospective) AUC**	**Location 1 (prospective) PR**	**Location 2 (external) AUC**	**Location 2 (external) PR**
Intracerebral haemorrhage	0.76	0.36	0.90	0.84
Ischaemic – all	0.85	0.85	0.87	0.88
Ischaemic – non-LVO	0.82	0.86	0.85	0.89
Ischaemic – LVO	0.86	0.82	0.91	0.87
Ischaemic - Thrombolysis	0.73	0.70	0.86	0.86
Ischaemic – Endovascular thrombectomy	0.85	0.80	0.87	0.75

AUC = area under the receiver operator curve, PR = Precision-Recall, LVO = large vessel occlusion

Chapter 5 - Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study, *Internal Medicine Journal*

Citation

Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, ‘Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study, *Internal Medicine Journal*, https://doi.org/10.1111/imj.15678

Statement of Authorship

Title of Paper	Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, ‘Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study, *Internal Medicine Journal*, https://doi.org/10.1111/imj.15678

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.
Overall percentage (%)	80%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.

Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate’s stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate’s stated contribution.

Name of Co-Author	Dr Sam Gluck		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Automated information extraction may be able to assist with the collection of stroke key performance indicators (KPIs). The feasibility of using natural language processing for classification-based KPI and datetime field extraction was assessed. Using free-text discharge summaries, random forest models achieved high levels of performance in classification tasks (AUC 0.95-1.00). The datetime field extraction method was successful in 29/43 (67.4%) cases. Further studies are indicated.

Manuscript

Introduction

In stroke, the adherence to selected key performance indicators (KPIs) is associated with lower risk of disability and death.[1] The "Acute Stroke Clinical Care Standard" produced by the *Australian Commission on Safety and Quality in Health Care* outlines seven quality statements with associated KPIs for Australian stroke centers.[2] The routine measurement and recording of these KPIs can be time consuming. Automated information extraction with natural language processing[3] may be able to assist with the recording of these KPIs.

This pilot study aims to assess the feasibility of using natural language processing to automatically extract information regarding the Australian national KPIs from free-text stroke discharge summaries. It included (1) classification tasks, with the primary outcome of the area under the receiver operator curve (AUC) for the identification of whether an individual had received thrombolysis on-site, and (2) the extraction of datetime information.

Method

Data collection

Consecutive individuals admitted to the Royal Adelaide Hospital Stroke Unit in the eight months following 1/1/2020, with a discharge summary in the current electronic health record (integration in February-March 2020), were included in the study. The free-text synopses of these discharge summaries, which were written starting from standardised free-text templates, were collected for analysis. Data regarding KPIs was collected from existing departmental databases. This KPI data had been entered by medical and nursing staff and provided the gold-standard for comparison. Individuals with missing outcome data were excluded from analysis for that individual outcome.

Selection of KPIs

Five binary datapoints were selected that would be required to calculate KPIs from the "Acute Stroke Clinical Care Standard". These datapoints included: the type of stroke (ischaemic vs intracerebral haemorrhage), whether an individual had atrial fibrillation/flutter, and whether they received thrombolysis on-site, endovascular thrombectomy, and/or an anticoagulant on discharge (irrespective of reason indication, or who had a documented contraindication). These datapoints were selected such that the datasets would not be so unbalanced as to prevent meaningful interpretation of the results.

Classification tasks

Initially, negation detection (see Supplementary Information 1 for Glossary) was applied to the free-text discharge summaries. Subsequently punctuation and stopwords were removed. Prior to use in classification experiments, count vectorisation was used to transform the text. This process included the conversion of text into unigrams, bigrams and trigrams.

Prior to classification experiments the dataset was split into training and testing datasets (train/test split 80%/20%). Models were developed on the training dataset using 5-fold cross-validation. These models included logistic regression, decision tree, random forest algorithms and artificial neural networks.

The best performing model was applied to the unseen test data for all five selected datapoints, and area under the receiver operator curve (AUC) calculated. The cut-off scores in the analysis of the test dataset to generate binary classifications were derived using Youden's index on the training dataset. To demonstrate the proportion of cases in which the algorithm had a high level of certainty, two additional cut-off scores were employed for the primary outcome (identification of on-site thrombolysis): >99% positive predictive value (PPV), and >99% negative predictive value (NPV).

Datetime extraction

Datetime extraction was conducted only for on-site thrombolysis. Prior to extraction with regular expression operations, text from the discharge summary synopsis was made lower case and split into separate sentences. These sentences were searched for a pre-defined set of keywords ("thrombolysis", "thrombolysed", "tenecteplase" and "alteplase"). The first sentence in which one of these key terms appeared was searched for words with formatting similar to a date or time with regular expression operations. The earliest datetime located in this sentence was then extracted and compared to that which was previously recorded in the departmental database. Instances in which the automatically extracted and previously recorded datetimes differed by more than 5 minutes, or where no datetime could be extracted, were manually inspected to determine the likely reason for the discrepancy.

Statistical analysis

Pre-processing, classification experiments and statistical analysis were performed with open-source Python libraries including NLTK (including negation detection), SciKit Learn and Tensorflow.

Statement of ethics

The Central Adelaide Local Health Network Research Ethics Committee granted approval for this project.

Results

Patient characteristics

438 individuals were included in the study. This cohort included 375 (85.6%) individuals with ischaemic strokes and 63 (14.4%) with intracerebral haemorrhage. Atrial fibrillation or flutter was recorded in 130 (29.7%) individuals. 43 patients received thrombolysis on-site (9.8% of total stroke, 11.5% of ischaemic stroke), and 103 (23.5% of total stroke, 27.5% of ischaemic stroke) received endovascular thrombectomy. The only outcome for which individuals were excluded due to missing outcome data was anticoagulation on discharge. The number of patients discharged with an anticoagulant or who had a documented contraindication was 169, with 176 discharged without an anticoagulant, and 93 without specification with respect to anticoagulants (and were therefore not included in the analysis of this outcome).

Classification tasks

The best performing model on the training dataset was the random forest model with 1,000 decision trees and employing the entropy criterion. When random forest models were applied to the test dataset, high levels of performance were achieved for identifying the type of stroke (AUC 1.00, accuracy 1.00), whether an individual received thrombolysis on-site (AUC 0.97, accuracy 0.93), whether an individual received endovascular thrombectomy (AUC 1.00, accuracy 1.00), and whether an individual had atrial fibrillation/flutter (AUC 0.97, accuracy 0.91) (see Table 1). Slightly lower performance was returned for the identification of those on anticoagulation on discharge (AUC 0.95, accuracy 0.91).

Examples in which the algorithm made a misclassification error regarding whether thrombolysis had been administered on-site included cases in which 3/6 underwent endovascular thrombectomy but did not receive thrombolysis, and 1/6 had a haemorrhagic stroke. 2/6 received thrombolysis, but at a different centre prior to transfer. In each of these cases the misclassification was a false positive with respect to whether thrombolysis had been administered on-site.

When the two additional high-certainty cut-off scores were applied to the test dataset (n=88) for the primary outcome (the identification of on-site thrombolysis), 6/88 (6.8%) individuals were above the high-PPV cut-off score (high-certainty positives). 69/88 (78.4%) individuals were below the high-NPV cut-off score (high-certainty negatives). The number of individuals between the two cut-off scores was 13/88 (14.8%).

Datetime extraction

Of the 43 cases that received thrombolysis on-site, the automated extraction method successfully retrieved the date and time in 23/43 (53.4%) cases. Of the remaining cases in which the automatically and human extracted date and time did not match, in 10/43 (23.3%) cases the time was able to be successfully extracted, but not the date. In 6/43 (14.0%) instances the automated method provided a datetime that was closer to that reported in the discharge summary than that which had been recorded in the database. Other cases, 4/43 (9.3%), included instances of thrombolysis with dates and times recorded in non-standard formats.

Discussion

This pilot study has provided evidence to support that accurate automated information extraction for stroke KPIs may be feasible from free-text discharge summaries. The performance seen in this study had some variability between data points. The use of high-certainty cut-off scores may facilitate automated information extraction, although human data extraction would still be required for indeterminant cases.

The high levels of classification performance achieved for most data points in this study are perhaps not surprising given the information should be explicitly contained within the document being analysed. The lower performance in the identification of those with anticoagulation on discharge is likely due to this information not necessarily being explicitly included in the free-text portion of the discharge summary. This performance could be improved by using additional fields from the discharge summary (such as the discharge medication list, which is separate to the free-text synopsis used in this study) or additional document types (such as the discharge note written by the ward pharmacist). Ultimately, even if accuracy were persistently unacceptable in some tasks despite use of additional fields and note-types, automating some, but not all, data fields could still provide a degree of gain in efficiency.

The methods investigated in this study may be useful for other tasks, beyond the recording of KPIs. For example, automated information extraction may be able to help collect data for research registries, as registries collect data in addition to standard KPIs based on local priorities.[4] Automated information extraction may also be able to assist with the implementation of machine learning models.

The usefulness and performance of natural language processing based upon medical free-text is inherently dependent on the comprehensiveness of that medical documentation. Performance may also be higher if there is more standardised means of recording data such as datetimes. Accordingly, performance may differ at different institutions. The unbalanced nature of the training and test datasets is a relevant consideration (for example 85.6% ischaemic stroke as compared to 14.4% with intracerebral haemorrhage), particularly when collecting information for KPIs. Where KPIs are met at a rate nearing 100% of individuals, it becomes difficult to assess the performance of machine learning algorithms, and may artificially make the performance appear higher than it would have been if the most prevalent category comprised a smaller proportion of the dataset. It is also possible that unrecognised errors may occur in the original manual recording of information in training and testing datasets, and that such errors may influence model development and performance evaluation.

Future research in this area could aim to derive models from larger datasets and to validate these models on prospective and external datasets. Similar models could also be investigated in the automated information extraction of other stroke registry or research data (such as for use in other machine learning algorithms). The methods investigated in this study may also be applied to the automated information extraction of KPI and research data in other medical specialties.

Conflicts of Interest: The authors declare they have no conflict of interest.

Funding: No funds, grants, or other support was received.

References

1 Urimubenshi G, Langhorne P, Cadilhac DA, Kagwiza JN, Wu O. Association between patient outcomes and key performance indicators of stroke care quality: A systematic review and meta-analysis. *Eur Stroke J*. 2017; **2**: 287-307.

2 Australian Commission on Safety and Quality in Health Care. Acute Stroke Clinical Care Standard 2019.

3 Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*. 2021; **38**: 4-9.

4 Cadilhac DA, Kim J, Lannin NA, Kapral MK, Schwamm LH, Dennis MS*, et al.* National stroke registries for monitoring and improving the quality of hospital care: A systematic review. *Int J Stroke*. 2016; **11**: 28-40.

Table 1: Results of binary classification tasks in automated information extraction from free-text discharge summaries

Content	AUC	PR score	TP	FN	TN	FP	Sensitivity	Specificity	PPV	NPV	F1 Score	Accuracy
What type of stroke?	1.00	1.00	74	0	14	0	1.00	1.00	1.00	1.00	1.00	1.00
Was endovascular therapy given?	1.00	1.00	21	0	67	0	1.00	1.00	1.00	1.00	1.00	1.00
Was thrombolysis given on-site?	0.97	0.85	13	0	69	6	1.00	0.92	0.68	1.00	0.81	0.93
Was atrial fibrillation or flutter present?	0.97	0.94	26	1	54	7	0.96	0.89	0.79	0.98	0.87	0.91
Discharged on anticoagulant?	0.95	0.92	28	2	35	4	0.93	0.90	0.88	0.95	0.90	0.91

AUC = area under the receiver operator curve, PR score = Precision-Recall score (See Supplementary Information 1 for definition), TP = True positives, FN = false negatives, TN = true negatives, FP = false positives, PPV = positive predictive value, NPV = negative predictive value. Please see Supplementary Information 1 for definition of F1 score. Values of TP, FN, TN and FP are provided as raw numbers.

Supplementary Information 1 – **Glossary**

In the following definitions TP = true positive, FP = false positive, TN = true negative and FN = false negative.

Artificial neural network – A type of machine learning algorithm based upon the structure of neurons. These structures are composed of nodes (each with multiple inputs, a bias and activation function) that are connected by weights. These networks are classically trained through back-propagation, whereby a cost function is calculated after a prediction is made and then that cost function is used to update the weights connecting the nodes, prior to the generation of another prediction (with the aim of minimising that cost function).

Count vectorisation – A process through which a type of input is converted into a table of token (often word) frequency counts. For example, a passage of text may be converted into a table of counts of each of the individual words that appeared in that passage.

Decision tree – A type of machine learning algorithm that involves the development of multiple rules to divide data into segments of increasingly smaller size. This process results in a series of rules, which when applied sequentially can be represented as a tree of sequential decision points.

Entropy criterion – Entropy is a measure of the level of information impurity. The entropy criterion is a rule that can be applied when developing a decision tree algorithm, which means that splits in the decision tree will be selected so as to minimise information impurity and thereby optimise information gain.

F1 Score – An outcome metric, otherwise known as the F-score or F-measure, that is calculated as follows: F1-score = TP/(TP + 0.5(FP+FN)).

Logistic regression – A logistic regression model is one that applies a logistic function so that inputs (often numerical) can be used to predict a categorical (often binary) outcome. This task is achieved by predicting the probability of an individual class. Examples of logistic regression may include binomial, multinomial and ordinal models.

Negation detection – A process through which words following a negative term (such as “no” or “not”) are flagged as being distinct from the term should it not have followed such a negative term. For example, in the sentence “Since the patient had recently taken apixaban, they received endovascular thrombectomy, but not thrombolysis”, the word “thrombolysis” would be negated. One way in which a word may be flagged as negated is by adding a certain suffix so that it reads “thrombolysis_neg”.

Precision-Recall score – Precision is calculated as: Precision = TP/(TP+FP). Recall is calculated as: Recall = TP/(TP+FN). In a binary test, as the cut-off score is varied, precision and recall will also change. By varying a cut-off score through a range of thresholds, a precision-recall curve can be created. A precision-recall score, or average precision score, summarises a precision-recall curve by calculating the weighted mean of the precisions at each of a range of cut-off scores.

Python – A type of programming language, which can be used to write computer programs.

Random forest – A type of machine learning algorithm which utilises multiple decision trees.

Regular expression – A regular expression is a sequence of code that describes a particular search pattern. Regular expressions are present in many programming languages.

Stopwords – A collection of words which typically have grammatical purposes but add little in the way of substantive meaning to a sentence (such as “to”, “a”, and “the”).

Unigrams – A sequence of one item. By extension, bigrams are sequences of two contiguous items and trigrams are sequences of three contiguous items. In the context of natural language processing examples of a unigram, bigram and trigram respectively may include “cerebral”, “cerebral artery” and “middle cerebral artery”.

Youden's index – Otherwise known as Youden’s J statistic, Youden’s index is a statistic for a binary test that can be calculated as follows: Youden’s index = sensitivity + specificity – 1. By calculating this index for multiple possible cut-off scores, the cut-off score with the highest Youden’s index can be determined.

Chapter 6 - Improving the accuracy of stroke clinical coding with open-source software and natural language processing, *Journal of Clinical Neuroscience*

Citation

Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, 'Improving the accuracy of stroke clinical coding with open-source software and natural language processing', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.10.024

Statement of Authorship

Title of Paper	Improving the accuracy of stroke clinical coding with open-source software and natural language processing
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Gluck S, Koblar S, Jannes J & Kleinig T 2021, 'Improving the accuracy of stroke clinical coding with open-source software and natural language processing', *Journal of Clinical Neuroscience*, https://doi.org/10.1016/j.jocn.2021.10.024

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Sam Gluck
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.

Signature		Date	8/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Abstract

Clinical coding is an important task, which is required for accurate activity-based funding. Natural language processing may be able to assist with improving the efficiency and accuracy of clinical coding. The aims of this study were to explore the feasibility of using natural language processing for stroke hospital admissions, employed with open-source software libraries, to aid in the identification of potentially misclassified (1) category of Adjacent Diagnosis Related Groups (ADRG), (2) the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification (ICD-10-AM) diagnoses, and (3) Diagnosis Related Groups (DRG). Data was collected for consecutive individuals admitted to the Royal Adelaide Hospital Stroke Unit over a five-month period for misclassification identification analysis. 152 admissions were included in the study. Using free-text discharge summaries, a random forest classifier correctly identified two cases classified as B70 ("Stroke and Other Cerebrovascular Disorders") that should be classified as B02 (having received endovascular thrombectomy). A regular expression-based analysis correctly identified 33 cases in which ataxia was present but was not coded. Two cases were identified that should have been classified as B70D, rather than B70A/B/C, based on transfer to another centre within five days of admission. A variety of techniques may be useful to help identify misclassifications in ADRG, ICD-10-AM and DRG codes. Such techniques can be implemented with open-source software libraries, and may have significant financial implications. Future studies may seek to apply open-source software libraries to the identification of misclassifications of all ICD-10-AM diagnoses in stroke patients.

Manuscript

Introduction

Accurate clinical coding is integral to the activity-based funding of hospitals in Australia and overseas. Factors that make this process challenging include high patient numbers, resource limitations and medical documentation that can, at times, lack in the specific details required for this clinical coding task [1, 2]. The application of natural language processing to this task in stroke may be beneficial.

In acute public hospital admissions, activity-based funding allocation is functionally completed through a series of tiers of classification, which come under the heading of Australian Refined Diagnosis Related Groups (AR-DRG, version 10). Initially, a primary diagnosis/procedure determines the category of the Adjacent Diagnosis Related Groups (ADRG) that will be applied. For example, an ischaemic stroke which undergoes endovascular thrombectomy would be categorised as B02, whereas an ischaemic stroke that does not have this procedure would usually be categorised as a B70. Subsequently, additional diagnoses that were encountered during the admission (as outlined in the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification - ICD-10-AM) are attributed to a given admission. Examples of such ICD-10-AM diagnoses include “hemiplegia unspecified” and “dysarthria and anarthria”. Depending on the primary diagnosis/procedure, these ICD-10-AM diagnoses add variable levels of complexity (as quantified by the diagnosis complexity level - DCL) to ultimately provide an episode clinical complexity score (ECCS). The number of possible ECCS levels varies by ADRG, but typically involves up to three categories. It is this ECCS, in conjunction with the ADRG, that provides the ultimate Diagnosis Related Groups (DRG) subclassification. For example, within the B70 ADRG (“Stroke and Other Cerebrovascular

Disorders"), different ECCS distinguish the B70A DRG ("Stroke and Other Cerebrovascular Disorders, Major Complexity") from B70B DRG ("Stroke and Other Cerebrovascular Disorders, Intermediate Complexity") subclassification. This final DRG, in conjunction with factors such as length of stay, hospital acquired complications, avoidable hospital readmissions, in conjunction with the national efficient price, influence hospital activity-based funding [3].

The level of interpretation that can be exercised by a clinical coder is limited. Accordingly, specific phrasing in medical documentation is important to aid with accurate clinical coding [1]. Demands on medical staff may, at times, limit the ability to include such specificity in documentation. Therefore, additional means by which to assist doctors and clinical coders to facilitate this process efficiently and accurately is an ongoing need.

Natural language processing involves the application of computers to human text, and has previously been demonstrated to be successful in providing meaningful classifications of medical text [4]. Therefore, it seems feasible that natural language processing may assist with the classifications involved in clinical coding. Such methods have been investigated and shown promise in multiple areas [5], although such research specifically in stroke has been limited. Additionally, with the increasing availability of open-source software libraries, if such libraries prove useful in this task, it may increase the accessibility of these techniques to a broader group of healthcare institutions.

The aims of this study were to explore the feasibility of using natural language processing, employed with open-source software libraries, to aid in the identification of potentially misclassified: (1) category of ADRG (primary diagnosis/procedure, namely B02 vs B70), (2) ICD-10-AM diagnoses (the presence or absence of ataxia), and (3) DRG (that is not solely based upon ICD-10-AM diagnoses, namely B70D vs B70A/B/C), for a given stroke hospital admission. These aims were selected to encompass several different methods of analysis that could be employed.

Materials and methods

Data collection

For misclassification identification analysis, data was collected on consecutive individuals admitted to the Royal Adelaide Hospital Stroke Unit between 1/9/2020 to 1/2/2021. Data was collected including ADRG classification, DRG classification, ICD-10-AM diagnoses, length of stay, discharge destination and free-text medical discharge summaries. This data was collected from existing institutional and departmental databases that are maintained by administrative and neurology staff. The clinical coding that had been applied to this dataset was manually reviewed by the investigators with individual case note reviews.

Previously collected data from a non-overlapping period (1/1/2020 to 31/8/2020) were used in the development of classification strategies and models. It should be noted that version 10 of the AR-DRG was implemented on 1/7/2020.

Data pre-processing

Free-text medical discharge summaries underwent several pre-processing steps prior to analysis. This included the application of negation detection, removal of capitalisation, and stopword removal. Stemming and lemmetization were not performed. Prior to use in random forest classifiers, text underwent count vectorization. All pre-processing and analyses were conducted with open-source Python libraries (namely NLTK and SciKit Learn) [6, 7].

ADRG classification (B02 vs B70) – Random forest classifier

The identification of potentially misclassified ADRGs was performed using a previously derived random forest classifier for the identification of whether endovascular thrombectomy had occurred [8]. This model was derived using the non-overlapping dataset from the stroke unit (individuals admitted in the first eight months of 2020). The classifier included 1,000 decision trees and used the entropy criterion. This random forest classifier was applied to determine its classification accuracy. As an alternative approach, rather than employing an individual cut-off score, probabilities were generated that each given case should have been classified as B02. After the generation of these probabilities, cases previously classified as B70 that had been allocated the highest probability of truly having an ADRG of B02 underwent a case-note review to determine their true ADRG. This process was repeated until the first previously-classified B70 case was confirmed to have an ADRG of B70 accurately allocated.

ICD-10-AM classification (ataxia present or absent) – Regular expression-based analysis

The ICD-10-AM diagnosis of “ataxia unclassified” (R270) was randomly selected using a random number generator as the ICD-10-AM diagnosis for analysis. Logistic regression was employed on cases from the non-overlapping dataset to identify terms with the most strongly associated coefficients associated with the R270 (and “ataxic gait” - R260) label. Subsequently, two partial terms “ataxi” and “coord” were employed in a regular expression-based analysis whereby the presence of these terms (in a non-negated fashion) was considered to represent likely presence of ataxia.

DRG classification (B70D vs B70A/B/C) – Datetime and other existing fields

Given that DRG B70A/B/C distinctions are based upon the allocation of ICD-10-AM diagnoses, the distinction between these categories was not further interrogated (see above regarding ICD-10-AM diagnosis analysis). The distinction between B70D and B70A/B/C is based upon a non-endovascular thrombectomy stroke being transferred to another centre and having a length of stay of less than five days [9]. Existing admission and discharge datetime fields were parsed and used to calculate length of stay. Discharge destination, as collected from existing departmental databases, was dichotomised to transfer vs non-transfer.

Outcome analysis

The primary outcome was number of misclassified B02 ADRG cases identified using natural language processing techniques. Performance metrics including accuracy, F1 score, sensitivity, specificity, positive predictive value and negative predictive value were calculated for the machine learning models for each of the aims. In aims (1) and (3) where ARDG or DRG classifications were found to be misclassified, the current national weighted activity unit (NWAU) calculator was used to produce an estimate of the potential financial difference such a misclassification may have for the health-system [10].

Ethical approval

This study received ethical approval from the Central Adelaide Local Health Network Research Ethics Committee.

Results

Patient characteristics

The total number of patients included in the misclassification analysis dataset, admitted between 1/9/2020 to 1/2/2021, was 152. In this group, the mean age of the patients was 74.6 years-old (standard deviation 13.2 years, median 76.6 years-old, IQR 67.2 years – 84.2 years). 74 patients were female (48.7 %). Regarding the original coding of the DRGs, B70C was the most common classification (46 cases), followed by B70B (44 cases), B70A (25 cases), B70D (16 cases), B02B (10 cases), B02A (6 cases) and B02C (5 cases). The most frequently coded ICD-10-AM diagnoses were G819 hemiplegia unspecified (67), and R471 dysarthria and anarthria (63). R270 ataxia unspecified was present in 19 cases and R260 ataxic gait in 5 individuals. Following case note review: two B70 cases were reclassified as B02; there were 33 additional cases in which ataxia was identified and one in which ataxia had been labelled as present that was absent; and two cases (one B70B and one B70C) that were reclassified as B70D.

ADRG classification (B02 vs B70) – Random forest classifier

When the random forest classifier was applied, it was found to have an accuracy of 0.974 and F1 score of 0.905 (see Table 1). The two previously classified B70 cases allocated the highest probability of being B02 by the random forest classifier had both in fact undergone endovascular thrombectomy. The third highest probability previously classified B70 was correctly classified. Using the current NWAU calculator the estimated difference in price for the two identified cases was positive $18,369 and $15,668 respectively.

ICD-10-AM classification (ataxia present or absent) – Regular expression-based analysis

The application of the regular expression-based analysis had an accuracy of 0.967 and F1 score of 0.955 (see Table 1). 35 cases were identified by the regular expression analysis as likely having ataxia, but not having been coded as R270 or R260. On case note review, of these identified cases, 33 cases did in fact have ataxia (94.3%). The two cases in which the analysis incorrectly attributed ataxia included a case where ataxia appeared as a possible feature of the past medical history and a case in which the phrase “normal coordination” was used (that was not tagged by the negation detection methods employed).

DRG classification (B70D vs B70A/B/C) – Datetime and other fields

When this datetime and other field analysis was applied there was an accuracy of 1.0 and F1 score of 1.0. Two cases were highlighted by the analysis as having been misclassified. These cases included one B70B and one B70C that were transferred to other centres in fewer than five days and accordingly would be classified as B70D. The estimated price difference for the two identified cases was negative $7,474 and $2,589.

Discussion

This study has demonstrated that a variety of techniques may be useful to help identify misclassifications in ADRG, ICD-10-AM and DRG codes. These techniques include machine learning classifiers, regular expression-based analyses and the use of existing data fields. Such techniques can be implemented with open-source software libraries in the stroke inpatient setting, and have significant financial implications.

In this study, an estimate of probability from a random forest classifier was used to stratify B70 cases by likelihood of misclassification. An alternative approach would be to set a cut-off score for a machine learning classifier such that it makes estimates of most likely category. In that case, instances in which the expected category and actual category do not align could then be examined.

It should be noted that all medical-text analyses performed in this study were conducted using only the free-text medical discharge summary. The performance of such methods is inherently dependent on the quality and thoroughness of the documentation being analysed and accordingly may differ between sites. It is likely that the use of additional forms of documentation (such as ward round notes and prescribing records), may improve the performance of some misclassification analyses.

The quality of the data entered in electronic health records may be a challenge for such computer-assisted strategies. Potential areas in which quality of data entry could be improved have previously been identified to include variability in diagnosis description, high quota expectations, staffing and budget restraints, and a lack of clear chart documentation [11-13]. Previous computer-assisted clinical coding studies have examined strategies including providing contextual information, negation detection and collaboration between coding supervisors, information technology coordinators and physicians to help improve the performance of computer-assisted clinical coding despite these limitations [5, 14-16].

As a pilot study this project has several limitations. This study was performed at a single centre. Additionally, English-language only medical text was analysed. It is possible that not all ADRG, ICD-10-AM, and DRG classifications will demonstrate the performance demonstrated for those selected in this study. Additionally, as new versions of the ADRG are produced, this may affect future performance. It is also worth noting the limitations of the algorithms employed in this study given that random forest algorithms are relatively computationally taxing and have more limited interpretability than other algorithms (such as individual decision trees).

Future research in this area may seek to employ similar techniques in other aspects of clinical coding in stroke. For example, such studies may seek to apply open-source software libraries to the identification of misclassifications of all ICD-10-AM diagnoses in stroke patients. If such models were successfully derived and validated, implementation studies would be warranted, examining whether their use could help improve clinical coding efficiency and accuracy.

Conflict of Interest

The authors declare that they have no conflict of interest.

Sources of support

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Hay P, Wilton K, Barker J, Mortley J, Cumerlato M. The importance of clinical documentation improvement for Australian hospitals. Health Inf Manag. 2020;49:69-73.

[2] Shepheard J. Clinical coding and the quality and integrity of health data. Health Inf Manag. 2020;49:3-4.

[3] Independent Hospital Pricing Authority. Australian Refined Diagnosis Related Groups (AR-DRGs). 2020.

[4] Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep Learning Natural Language Processing Successfully Predicts the Cerebrovascular Cause of Transient Ischemic Attack-Like Presentations. Stroke. 2019;50:758-60.

[5] Campbell S, Giadresco K. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Inf Manag. 2020;49:5-18.

[6] Bird S, Klein E, Loper E. Natural Language Processing with Python: O'Reilly Media Inc.; 2009.

[7] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825−30.

[8] Bacchi S, Gluck S, Koblar S, Jannes J, Kleinig T. Automated Information Extraction from Free-Text Medical Documents for Stroke Key Performance Indicators: A Pilot Study. In submission. 2021.

[9] Independent Hospital Pricing Authority. Australian Refined Diagnosis Related Groups Version 10.0. 2020.

[10] Independent Hospital Pricing Authority. National Weighted Activity Unit (NWAU) calculators. 2020.

[11] Alonso V, Santos JV, Pinto M, Ferreira J, Lema I, Lopes F, et al. Health records as the basis of clinical coding: Is the quality adequate? A qualitative study of medical coders' perceptions. Health Inf Manag. 2020;49:28-37.

[12] Doktorchik C, Lu M, Quan H, Ringham C, Eastwood C. A qualitative evaluation of clinically coded data quality from health information manager perspectives. Health Inf Manag. 2020;49:19-27.

[13] Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. BMC Health Serv Res. 2017;17:766.

[14] Perera S, Sheth A, Thirunarayan K, Nair S, Shah N. Challenges in understanding clinical notes. Proceedings of the 2013 international workshop on Data management & analytics for healthcare - DARE '132013. p. 21-6.

[15] Rinkle V. Computer assisted coding - a strong ally, not a miracle aid. Journal of Health Care Compliance. 2015;17:55–8.

[16] Whittle K. Winning the coding trifecta: CAC, CDI, and ICD-10. Journal of AHIMA. 2016;87:28–31.

Table 1: Results of the application of open-source software to stroke clinical coding

Method of identification	**Clinical coding task**	**True positive**	**False negative**	**True negative**	**False positive**	**Sensitivity (%)**	**Specificity (%)**	**Positive predictive value (%)**	**Negative predictive value (%)**	**F1 Score**	**Accuracy**
Random forest classifier	ADRG: B02 vs B70	19	4	129	0	82.6	100	100	97	0.905	0.974
Regular expression-based analysis	ICD-10-AM: Ataxia present or not	53	3	94	2	94.6	97.9	96.4	96.9	0.955	0.967
Datetime and other field analysis	DRG: B70D vs B70A/B/C	18	0	134	0	100	100	100	100	1	1

ADRG = Adjacent Diagnosis Related Groups, IDC-10-AM = International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification, DRG = Diagnosis Related Groups.

Chapter 7 - Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study, *Internal and Emergency Medicine*

Citation

Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study', *Intern Emerg Med*, https://doi.org/10.1007/s11739-019-02265-3

Statement of Authorship

Title of Paper	Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Menon D, Jannes J, Kleinig T & Koblar S 2020, 'Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study', *Intern Emerg Med*, https://doi.org/10.1007/s11739-019-02265-3

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Sam Gluck
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.

Signature		Date	8/1/2022

Name of Co-Author	Dr Yiran Tan		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Ivana Chim		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Joy Cheng		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Toby Gilbert		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Prof David Menon		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	4/3/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.

Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Length of stay (LOS) and discharge destination predictions are key parts of the discharge planning process for General Medical hospital inpatients. It is possible that machine learning, using natural language processing, may be able to assist with accurate LOS and discharge destination prediction for this patient group. Emergency Department triage and doctor notes were retrospectively collected on consecutive General Medical and Acute Medical Unit admissions to a single tertiary hospital from a two-month period in 2019. This data was used to assess the feasibility of predicting LOS and discharge destination using natural language processing and a variety of machine learning models. 313 patients were included in the study. The artificial neural network achieved the highest accuracy on the primary outcome of predicting whether a patient would remain in hospital for >2 days (accuracy 0.82, area under the received operator curve 0.75, sensitivity 0.47, specificity 0.97). When predicting LOS as an exact number of days, the artificial neural network achieved a mean absolute error of 2.9 and a mean squared error of 16.8 on the test set. For the prediction of home as a discharge destination (vs any non-home alternative), all models performed similarly with an accuracy of approximately 0.74. This study supports the feasibility of using natural language processing to predict General Medical inpatient LOS and discharge destination. Further research is indicated with larger, more detailed, datasets from multiple centres to optimise and examine the accuracy that may be achieved with such predictions.

Manuscript

Introduction

Discharge planning is the process by which an individualized plan is made for a patient to leave hospital and receive ongoing support in the community, preventing readmission [1]. Effective discharge planning can reduce risk of readmission, improve patient satisfaction and reduce length of stay (LOS) [2]. Discharge planning may involve multiple components including input from allied health staff, discussions with community healthcare providers, estimating and communicating an estimated discharge date (EDD) and discharge destination. Accurate prediction of EDD, based on predicted LOS, contributes to better preparation and allocation of resources, improved patient safety and satisfaction, and reduced healthcare costs [3, 4].

The generation of an EDD, based upon the predicted LOS, and nominating appropriate discharge destinations are important parts of the discharge planning process [5]. However, predicting EDD and discharge destination at the time of admission can be challenging, particularly when fewer senior medical staff are performing admitting roles. Numerous patient and hospital variables may influence LOS. Patient variables that have previously been found to influence LOS include age, ethnic group, marital status and source of referral [6, 7].

It is possible that machine learning (ML), in particular deep learning (DL), may be able to enhance LOS prediction accuracy. ML encompasses a wide variety of algorithms, including logistic regression, random forest models, support vector machines and DL [8]. DL may be considered as a type of ML, which is characterized by the use of artificial neural networks. Such methods have previously been applied to general medicine patients in the prediction of outcomes such as in-hospital mortality [9]. ML has also previously been applied to LOS prediction in a surgical setting, particularly with regards to neurosurgical patients [10, 11]. When ML is applied to human language (either in a written or spoken format) this is referred to as natural language processing (NLP) [12]. Since some variables influencing LOS are likely described in free-text medical notes and would not be routinely entered in a discrete data field (e.g. administrative data fields) in an electronic health record [13], such as descriptive aspects of a patient's clinical presentation, natural language processing may improve the accuracy of LOS estimates. The application of natural language processing, in conjunction with DL, to free-text data specifically in the General (Internal) Medicine patient population has not been examined previously.

We therefore assessed the feasibility and initial accuracy of predicting the LOS (and thereby EDD) and discharge destination for a diverse array of General Medical patients using natural language processing.

Materials and Methods

Data collection and pre-processing

Data was retrospectively collected on consecutive patients admitted to the Acute Medical Unit (AMU), or General Medical Unit of a tertiary hospital (The Royal Adelaide Hospital) over the course of a two-month period in 2019. Data in the form of free-text that would have been available at the time of admission to hospital were collected. This data included demographic details, the Emergency Department (ED) triage note, the note by the ED doctor, and initial investigation results, if recorded in the ED doctor note. Each patient's record was reviewed to determine the LOS and discharge destination on separation. It should be noted that some patients would have come into hospital who were already from non-home destinations, such as residential care facilities. When discharged to a residential care facility, regardless of the patients' pre-hospital living circumstances, this was considered a "non-home" destination. All of the data collected that was used in the predictive models would have been available within approximately 4 hours of the patient's arrival to hospital, prior to transfer from the ED to an inpatient ward.

At the tertiary hospital at which the study was conducted the Acute Medical Unit (AMU) and General Medical Unit admit a wide variety of patients with pathological processes involving multiple organ systems. Patients who are expected to have a LOS that is less than or equal to two days are admitted to the AMU, with patients expected to have a longer LOS admitted to General Medical teams. The decision to admit to AMU lies with the on-call medical registrar. Similar short-stay medical units have been employed in other centres with a focus on managing patients with a brief predicted LOS [14].

Admission allocation to AMU or General Medicine provided a comparative human-estimated LOS (≤2 days or >2 days). Allocation was a dichotomous decision made by the doctor at the time of admission. Human estimates for discharge destination or LOS as a continuous outcome were unavailable, as these are not currently routinely estimated or recorded on admission.

Following collection, data were pre-processed prior to use in ML models. The type of pre-processing required differed according to the ML model that was to be applied. Initially, words were converted to word stems. For example, words including "improve", "improves", "improvement", "improving" and "improved" would all have been shortened to the word stem "improv". Prior to use with convolutional neural networks, free-text data were converted into arrays of sequences of numbers (tokenisation). In these arrays each number represented a unique word stem and the position of the word within the original free-text. All arrays were made an equal length through the addition of blank tokens to the end of shorter arrays (padding). Prior to use in other models (logistic regression and random forest) negation detection was applied. Negation detection involves the flagging of words following negating terms (for example in the phrase "not painful", "painful" would be flagged as negated). Stop words (e.g. “the” and “is”) and non-letter characters (e.g. punctuation) were removed. Following this process free-text data were converted into arrays with frequency counts of each unique word stem.

Data were then randomly split into a training dataset, and a hold-out test dataset. This split was performed once. The data were split 85%/15% between the training and testing datasets respectively. This proportion of train/test split was selected to maximise training data in the context of the relatively small sample size of a pilot study.

Classification experiments

The assessment of LOS using a dichotomous outcome (≤2 days or >2 days) and discharge destination by dichotomous outcome of home or non-home was examined using classification experiments. All ML models were developed using the training dataset. This development was conducted using 5-fold cross-validation. This model development process was conducted by starting with a simple model, with few layers and few nodes per layer. Sequential analyses were conducted, on each occasion adding further complexity, in the form of further layers and nodes per layer, until accuracy no longer improved, and the final model architecture was decided. Grid-search functions were also used to optimise ML model structures and hyperparameters. Once models were considered optimised on the training set (optimisation for accuracy), they were then tested on the hold-out test dataset. These methods are similar to those used in other medical ML studies, and have been described previously [15]. Youden’s index was used to calculate cut-off scores [16]. The predicted classifications based upon these cut-off scores were used to calculate classification metrics including sensitivity, specificity and accuracy. This process was conducted individually for four types of ML model: convolutional neural network, artificial neural network, logistic regression and random forest models. The primary outcome of the study is the accuracy of prediction for

LOS as a dichotomous outcome (less than or equal to two days, or greater than two days – as defined by the number of midnights spent in hospital, rather than a 48-hour cut-off). The accuracy of predictions for discharge destination as a dichotomous outcome (home vs non-home) was a secondary outcome.

Following the development and testing of the logistic regression model for assessing the dichotomous LOS outcome, word stems with the most strongly predictive coefficients were identified.

Regression experiments

The mean absolute error for predictions of LOS as a continuous outcome was assessed using regression experiments. For each of the DL models assessed (convolutional neural network and artificial neural network), development again occurred on the training set using 5-fold cross-validation and grid-search functions. Once models were considered optimised (for mean absolute error), they were then assessed on the hold-out test set. The mean absolute error (MAE) for predictions of LOS as a continuous outcome (number of days) was a secondary outcome of the study.

DL model structures

The optimised artificial neural network model structure that was applied in the regression and classification experiments was comprised of six fully-connected layers (see Figure 1). The number of nodes per layer were 512, 512, 256, 256, 256 and 32 respectively. The convolutional neural network model structure involved an embedding layer, a convolutional layer, a maximum pooling layer, and then five fully-connected layers (number of nodes 512, 256, 128, 64, 32), interspersed by dropout layers (with a dropout rate of 0.1).

Statistical analysis

Outcome metrics including accuracy, area under the receiver-operator curve (AUC) and mean absolute error were calculated using open source Python libraries (SciKit Learn).

Statement of ethics

Ethics approval was granted for this project by the Central Adelaide Local Health Network Research Ethics Committee (HREC/19/CALHN/209). A waiver of consent was acquired from the Ethics Committee.

Results

Patient characteristics

All 313 patients who were screened for the study were included (i.e. none were excluded). The mean age was 70.4 (SD 19.2) years-of-age, and 54.8% were female. The average LOS was 6.8 [IQR 2-9] days. The number of patients who stayed for two days or fewer was 94 (30.0%). The number of patients who were discharged home was 186 (59.4%).

Primary Outcome

Prediction of LOS - dichotomous outcome

When predicting whether a patient would remain in hospital for >2 days, the artificial neural network achieved the highest accuracy and the highest AUC (accuracy 0.82, AUC 0.75, sensitivity 0.47, specificity 0.97, PPV 0.88, NPV 0.81, F1 score 0.61) (see Table 1). Logistic regression and the convolutional neural network achieved similar accuracies (0.8 and 0.78 respectively) and AUC (0.68 and 0.66 respectively) (see Table 1). The random forest model achieved an accuracy of 0.72 and AUC 0.62. The accuracy achieved by the initial classification by the human medical registrar was 0.92 (see Table 1).

In the logistic regression model, the word stems that were found to be most predictive of staying in hospital >2 days included “last”, “per”, “fall”, “febril”, “home” and “ICU”. Word stems that were found to be most predictive of LOS ≤2 days included “state”, “15”, “improv”, “arriv”, “AMU” and “pain”.

Secondary outcomes

Prediction of LOS – continuous outcome

The artificial neural network achieved a mean absolute error of 2.9 and a mean squared error of 16.8 on the test set. The convolutional neural network model had similar scores with a mean absolute error of 3.1 and mean squared error 18.8.

Prediction of discharge destination

For the prediction of home as a discharge destination (vs any non-home alternative), all models performed similarly. Logistic regression, artificial neural network, convolutional neural network and random forest models achieved accuracies of 0.76, 0.74, 0.72 and 0.72 respectively (see Table 1). No medical registrar comparison was available for the prediction of discharge destination.

Discussion

The results of this study support the feasibility of using natural language processing DL to predict General Medical inpatient LOS and discharge destination, and indicate that further investigation is required.

These results support those of other studies that have shown that DL may have utility in predicting LOS and discharge destination. Surgical fields in which the method has been trialled include neurosurgery [11, 17-19], orthopaedics [10, 20, 21] and general surgery [22]. There have been fewer studies predicting LOS or discharge destination in medical specialties, namely in the fields of cardiology [23, 24] and thoracic medicine [25]. The accuracies of these previous studies have varied based upon the selection of outcome metric, sample size and model complexity. Models with AUC as high as 0.94 have been reported for selected patient populations [23]. The majority of such studies employ structured data fields for input data, as has been performed in the prediction of LOS of patients with heart failure or diabetes [26, 27]. For example, a field may be present for each of age, gender, smoking status (Y/N) and the presence of type 2 diabetes (Y/N). The findings of the presented study are significant in that natural language processing was applied to non-structured free-text input data (rather than structured/discrete data fields), and encompassed patients with a wide variety of illnesses and presenting issues.

It is reasonable to hypothesise that the combination of discrete/structured data and free-text data may result in better performance than either input-type alone (given that the free-text data may include more information than that available in the routinely collected structured data). In future studies, incorporation of both types of data into an individual model could be investigated to see whether this approach may improve performance. An alternative or complementary strategy may be to employ natural language processing to automatically populate structured/discrete data fields, from free-text data. Further research investigating such combinations of structured and unstructured data is required.

The natural language processing method used in this study is similar to that described in other medical studies involving natural language processing [28]. In particular, the pre-processing strategies employed are similar to those which have been described previously [29, 30]. The natural language processing method used in this study has been successfully applied to the classification of transient ischaemic attack-like presentations [31]. A convolutional neural network was employed, rather than other structures such as a recurrent neural network, due to the results of previous studies involving the classification of medical free-text [31].

In view of the limited sample size of this pilot study, the primary outcome used to assess the feasibility of the DL for this application was a dichotomised classification task (LOS ≤2 days or >2 days). This method of analysing LOS is similar to that which has been used in other studies [23]. Larger samples sizes would allow for greater accuracy in the prediction of a continuous primary outcome (i.e. actual estimate of number of days LOS). Similarly, discharge destination was assessed as a dichotomised classification task (home vs non-home), as opposed to a multiple class classification task (e.g. home vs residential care facility vs other hospital vs other healthcare facility vs death), to facilitate the assessment of feasibility in this pilot study. In future studies, the prediction of multiple class outcomes may improve the possible future utility of the developed models.

The ≤2 days or >2 days LOS categories were selected since local admission criteria specify that patients with an expected stay ≤2 days are admitted to the AMU, and those with a longer expected stay are admitted to General Medicine. This decision is made by a senior medical registrar, and provided a convenient comparative human estimate of LOS. In this study the medical registrar estimated LOS had a high accuracy of 0.92. However, it is highly likely that the prophecy of a ≤2 day LOS indicated by allocation to the AMU is self-fulfilling, as the AMU has different staffing, resources and discharge expectations compared with a General Medical unit.

This study had several additional limitations including a small sample size and being conducted at a single site. Further, given the pilot nature of the project, no statistical significance tests were conducted to determine if one model was superior to another. Pre-hospital living circumstances are not routinely recorded and accordingly the number of patients already living in residential care facilities at the time of admission is uncertain. It should also be noted that the investigation (laboratory and radiology) results that were included in the input data for the ML algorithm were limited to those that the Emergency Department doctors had typed in their notes. Therefore, not all investigation results available at the time of admission were included in the analysis. Our current models would have a discharge prediction after approximately 4 hours (Emergency Department LOS). However, it would be unlikely to be detrimental to discharge planning if prediction availability was at 24 hours. Making a LOS prediction at 24 hours would substantially increase the available data for the models to use, which in turn may increase their accuracy.

Further research in this area may seek to investigate the accuracy that can be achieved in LOS and discharge destination prediction using larger datasets and datasets from multiple centres. Future machine learning models will ideally be compared against a prospectively provided prediction of LOS and discharge destination by admitting staff, in a general medical model less likely to be contaminated by self-fulfilling aspects of clinical care, such as different staffing and discharge expectations. The addition of image data and structured field data to that of the free-text data may also be investigated as a means to improve the accuracy of predictions. Further studies may also seek to prospectively validate and then assess the effects of the implementation of such models on a healthcare system.

Conclusion

Based on data from this single centre pilot study it appears feasible to use natural language processing to predict hospital length of stay and discharge destination of general medical inpatients based on written data available at the time of hospital admission. Further larger, multicentre, studies are required to investigate the significance of these findings.

Conflict of interest**:** The authors have no conflicts of interest to declare.

Funding source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. SG is supported by a Royal Adelaide Hospital Research Committee A.R Clarkson Scholarship.

References

1. Lin C-J, Cheng S-J, Shih S-C, Chu C-H, Tjung J-J (2012) Discharge Planning. International Journal of Gerontology 6 (4):237-240. doi:10.1016/j.ijge.2012.05.001

2. Goncalves-Bradley DC, Lannin NA, Clemson LM, Cameron ID, Shepperd S (2016) Discharge planning from hospital. Cochrane Database Syst Rev (1):CD000313. doi:10.1002/14651858.CD000313.pub5

3. Daghistani T, Elshawi R, Sakr S, Ahmed A, Al-Thwayee A, Al-Mallah M (2019) Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. International Journal of Cardiology. doi:https://doi.org/10.1016/j.ijcard.2019.01.046

4. Maharlou H, Kalhori S, Shahbazi S, Ravangard R (2018) Predicting Length of Stay in Intensive Care Units after Cardiac Surgery: Comparison of Artificial Neural Networks and Adaptive Neuro-fuzzy System. Healthcare Informatics Research 24 (2):109-117

5. Waring J, Marshall F, Bishop S, Sahota O, Walker M, Currie G, Fisher R, Avery T (2014). In: An ethnographic study of knowledge sharing across the boundaries between care processes, services and organisations: the contributions to 'safe' hospital discharge. Health Services and Delivery Research. Southampton (UK). doi:10.3310/hsdr02290

6. Liu Y, Phillips M, Codde J (2001) Factors influencing patients' length of stay. Aust Health Rev 24 (2):63-70

7. Khosravizadeh O, Vatankhah S, Bastani P, Kalhor R, Alirezaei S, Doosty F (2016) Factors affecting length of stay in teaching hospitals of a middle-income country. Electron Physician 8 (10):3042-3047. doi:10.19082/3042

8. Deo R (2015) Machine Learning in Medicine. Circulation 123 (1920-30). doi:10.1161/CIRCULATIONAHA.115.001593

9. Schwartz N, Sakhnini A, Bisharat N (2018) Predictive modeling of inpatient mortality in departments of internal medicine. Intern Emerg Med 13 (2):205-211. doi:10.1007/s11739-017-1784-8

10. Elbattah M, Molloy O (2016) Using Machine Learning to Predict Length of Stay and Discharge Destination for Hip-Fracture Patients. SAI Intelligent Systems Conference

11. Mulestein W, Akagi D, Davies J, Chambless L (2018) Predicting Inpatient Length of Stay After Brain Tumor Surgery: Developing Machine Learning Ensembles to Improve Predictive Performance. Neurosurgery. doi:doi: 10.1093/neuros/nyy343. [Epub ahead of print]

12. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J (2019) A guide to deep learning in healthcare. Nat Med 25 (1):24-29. doi:10.1038/s41591-018-0316-z

13. Mazzali C, Duca P (2015) Use of administrative data in healthcare research. Intern Emerg Med 10 (4):517-524. doi:10.1007/s11739-015-1213-9

14. Yong TY, Li JY, Roberts S, Hakendorf P, Ben-Tovim DI, Thompson CH (2011) The selection of acute medical admissions for a short-stay unit. Intern Emerg Med 6 (4):321-327. doi:10.1007/s11739-010-0490-6

15. Sidey-Gibbons JAM, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19 (1):64. doi:10.1186/s12874-019-0681-4

16. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3:32-35

17. Chambless L, Thompson R, Weaver K, Morone P, Kallos J, Akagi D, Muhlestein W (2017) Using a Guided Machine Learning Ensemble Model to Predict Discharge Disposition following Meningioma Resection. Journal of Neurological Surgery Part B: Skull Base 79 (02):123-130. doi:10.1055/s-0037-1604393

18. Hollon T, Parikh A, Pandian B, Tarpeh J, Orringer D, Barkan A, McKean E, Sullivan S (2018) A machine learning approach to predict early outcomes after pituitary adenoma surgery. Neurosurg Focus 45 (5):E8

19. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, Smith TR, Arnaout O (2018) Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World Neurosurg 109:476-486 e471. doi:10.1016/j.wneu.2017.09.149

20. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, Patterson BM, Krebs VE (2019) Development and Validation of a Machine Learning Algorithm After Primary Total Hip Arthroplasty: Applications to Length of Stay and Payment Models. J Arthroplasty 34 (4):632-637. doi:10.1016/j.arth.2018.12.030

21. Navarro SM, Wang EY, Haeberle HS, Mont MA, Krebs VE, Patterson BM, Ramkumar PN (2018) Machine Learning and Primary Total Knee Arthroplasty: Patient Forecasting for a Patient-Specific Payment Model. J Arthroplasty 33 (12):3617-3623. doi:10.1016/j.arth.2018.08.028

22. Chuang M-T, Hu Y-h, Lo C-L (2018) Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. International Transactions in Operational Research 25 (1):75-90. doi:10.1111/itor.12298

23. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH (2019) Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. Int J Cardiol 288:140-147. doi:10.1016/j.ijcard.2019.01.046

24. Yakovlev A, Metsker O, Kovalchuk S, Bologova E (2018) Prediction of in-Hospital Mortality and Length of Stay in Acute Coronary Syndrome Patients Using Machine-Learning Methods. Journal of the American College of Cardiology 71 (11). doi:10.1016/s0735-1097(18)30783-6

25. Goto T, Camargo CA, Jr., Faridi MK, Yun BJ, Hasegawa K (2018) Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 36 (9):1650-1654. doi:10.1016/j.ajem.2018.06.062

26. Turgeman L, May JH, Sciulli R (2017) Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. Expert Systems with Applications 78:376-385. doi:10.1016/j.eswa.2017.02.023

27. Morton A, Marzban E, Giannoulis G, Patel A, Aparasu R, Kakadiaris IA (2014) A Comparison of Supervised Machine Learning Techniques for Predicting Short-Term In-Hospital Length of Stay among Diabetic Patients. Paper presented at the 2014 13th International Conference on Machine Learning and Applications,

28. Chary M, Parikh S, Manini AF, Boyer EW, Radeos M (2019) A Review of Natural Language Processing in Medical Education. West J Emerg Med 20 (1):78-86. doi:10.5811/westjem.2018.11.39725

29. Rajput A (2019) Natural Language Processing, Sentiment Analysis and Clinical Analytics. arXivorg. doi:arXiv:1902.00679

30. Wong A, Plasek JM, Montecalvo SP, Zhou L (2018) Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy 38 (8):822-841. doi:10.1002/phar.2151

31. Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J (2019) Deep Learning Natural Language Processing Successfully Predicts the Cerebrovascular Cause of Transient

Ischemic Attack-Like Presentations. Stroke 50 (3):758-760.

doi:10.1161/STROKEAHA.118.024124

Table 1: Results for machine learning models applied to dichotomous classification tasks regarding length of stay and discharge destination

Outcome	Model	AUC	TP[1]	FN	TN	FP	Sensitivity (Recall)	Specificity	PPV (Precision)	NPV	F1 Score	Accuracy
LOS dichotomous outcome (≤2 days vs >2 days)	Medical Registrar[2]	-	12	3	34	1	0.8	0.97	0.92	0.92	0.86	0.92
	ANN	0.75	7	8	34	1	0.47	0.97	0.88	0.81	0.61	0.82
	CNN	0.66	8	7	31	4	0.53	0.89	0.67	0.82	0.59	0.78
	Random Forest	0.62	7	8	29	6	0.47	0.83	0.54	0.78	0.50	0.72
	Logistic regression	0.68	6	9	34	1	0.40	0.97	0.86	0.79	0.55	0.80
Discharge destination dichotomous outcome (home vs non-home)	ANN	0.64	28	4	9	9	0.88	0.50	0.76	0.69	0.81	0.74
	CNN	0.64	26	6	10	8	0.81	0.56	0.76	0.63	0.79	0.72
	Random Forest	0.68	24	8	12	6	0.75	0.67	0.80	0.60	0.77	0.72
	Logistic regression	0.64	32	0	6	12	1.00	0.33	0.73	1.00	0.84	0.76

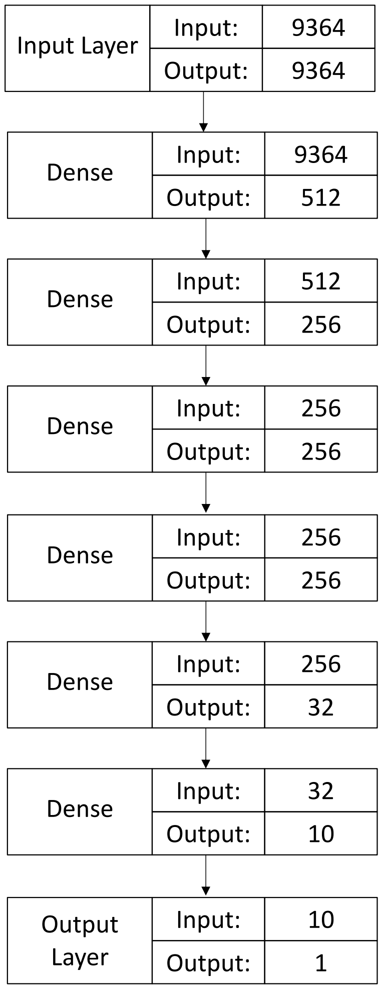
AUC = Area under the curve, ANN = artificial neural networks, CNN = convolutional neural networks, LOS = Length of Stay, TP = True positive, FN = False negative, TN = True negative, FP = False positive, PPV = Positive predictive value, NPV = Negative predictive value.

[1] In the section regarding LOS as a dichotomous outcome, true positives indicate individuals who stayed for ≤2 days, and the specified prediction model correctly predicted that the individual would stay for ≤2 days. In the section regarding discharge destination prediction,

true positives indicate the number of individuals who were discharged to home, and that the specified prediction model correctly predicted would be discharged to home.

[2] The “Medial Registrar” row in this table indicates the classification performance of the admitting medical registrar for an individual patient. This human estimated LOS was able to be derived because local admission criteria specify that patients with an expected stay ≤2 days are admitted to the Acute Medical Unit, and those with a longer expected stay are admitted to the General Medical Unit.

Figure 1: Diagram outlining the structure of the artificial neural network used in classification and regression experiments



Chapter 8 - Mixed-data Deep Learning in Repeated Predictions of General Medicine Length of Stay: A Derivation Study, *Internal and Emergency Medicine*

Citation

Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Jannes J, Kleinig T & Koblar S 2021, ‘Mixed-data Deep Learning in Repeated Predictions of General Medicine Length of Stay: A Derivation Study’, *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02697-w

Statement of Authorship

Title of Paper	Mixed-data Deep Learning in Repeated Predictions of General Medicine Length of Stay: A Derivation Study
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, Jannes J, Kleinig T & Koblar S 2021, 'Mixed-data Deep Learning in Repeated Predictions of General Medicine Length of Stay: A Derivation Study', *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02697-w

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Sam Gluck
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.

Signature		Date	8/1/2022

Name of Co-Author	Dr Yiran Tan		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Ivana Chim		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Joy Cheng		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Toby Gilbert		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.

Signature		Date	9/3/2022

Abstract

The accurate prediction of likely discharges and estimates of length of stay (LOS) aid in effective hospital administration and help to prevent access block. Machine learning (ML) may be able to help with these tasks. For consecutive patients admitted under General Medicine at the Royal Adelaide Hospital over an 8-month period, daily ward round notes and relevant discrete data fields were collected from the electronic medical record. These data were then split into training and testing sets (seven-month/one-month train/test split) prior to use in ML analyses aiming to predict discharge within the next two days, discharge within the next seven days and an estimated date of discharge (EDD). Artificial neural networks and logistic regression were effective at predicting discharge within 48 hours of a given ward round note. These models achieved an area under the receiver operator curve (AUC) of 0.80 and 0.78 respectively. Prediction of discharge within seven days of a given note was less accurate, with artificial neural network returning an AUC of 0.68 and logistic regression an AUC of 0.61. The generation of an exact EDD remains inaccurate. This study has shown that repeated estimates of LOS using daily ward round notes and mixed-data inputs are effective in the prediction of general medicine discharges in the next 48 hours. Further research may seek to prospectively and externally validate models for prediction of upcoming discharge, as well as combination human-ML approaches for generating EDDs.

Manuscript

Introduction

Awareness of the timing of likely discharges, accurate estimates of length of stay (LOS) and estimated dates of discharge (EDD) may help to manage hospital bed-state and avoid access block. Often these estimates are generated by staff in daily meetings, and then fed through to bed-flow coordinators via electronic medical records, emails, or further meetings. Consequently, such estimates may be delayed, which may be a barrier to effective discharges [1]. Machine learning (ML) may be able to help with the prediction of upcoming discharges and LOS in real-time and in an automated fashion.

Natural language processing applied to admission data has previously demonstrated moderate accuracy using artificial neural networks and logistic regression in the prediction of short general medicine LOS in a pilot study [2]. However, such one-off predictions at the time of admission will likely be limited in their ability to predict outcomes for patients once they have been admitted for several days. Research in other specialties has shown that repeated predictions of LOS, using new information gained during the hospital stay, can be used to refine LOS predictions throughout a hospital stay [3].

The aims of this study were to derive and assess the effectiveness of ML models using daily ward round notes to (A) predict the likelihood of discharge within the next two days (<48 hours), (B) predict the likelihood of discharge within the next seven days, and (C) generate an actual EDD, in a repeated fashion (on an ongoing daily basis).

Materials and Methods

Data collection and pre-processing

Ward round notes for patients admitted under, and discharged from, General Medicine at the Royal Adelaide Hospital over an 8-month period from 1/1/2020 were extracted from electronic medical records. Age, length of stay and the day of the week a note was written were also collected. Instances with missing data were excluded.

All text data underwent word stemming and removal of stopwords. Negation detection and count vectorisation was performed, with the formation of n-grams of 3 word stems in length. The maximum number of features was limited to 30,000. After pre-processing, data were split into training and testing sets based on date of admission (seven-month/one-month train/test split).

Model development and classification experiments

Logistic regression and artificial neural network models were developed on the training dataset using 5-fold cross-validation. In this process a basic model was employed first, for example an artificial neural network with few interconnected layers and nodes. Model complexity, in the form of additional layers and nodes, was increased in a progressive manner until the mean area under the receiver operator curve on the training set no longer improved. During this process different activation functions (including sigmoid and rectified linear unit activation functions) and hyperparameters were trialled.

After model structures and hyperparameters were optimised on the training dataset, models had their performance assessed on the unseen test dataset. This method was employed for the prediction of the primary outcome (discharge within the next two days from the time when the note was written) and the secondary outcome (discharge within the next seven days). The best performing model was then tested with high sensitivity and specificity cut-offs for the primary outcome.

After logistic regression classification experiments predicting discharge within the next two days, n-grams most and least associated with discharge within two days of note writing were extracted to aid with interpretability.

Regression experiments

Artificial neural network models were used in regression experiments aiming to predict LOS as an actual number of days from the time of the ward round note, and therefore provide an exact EDD. Similar to the classification experiments, the models were developed on the training dataset, prior to performance assessment on the test dataset.

Artificial neural network structure

The artificial neural network was built to incorporate both discrete data fields (current LOS, age, and day of week), as well as the text data (mixed-input data types). The artificial neural network structure that was used comprised two separate artificial neural networks that were concatenated into a single model (see Figure 1 for a schematic of the artificial neural network and the "Results" section for details regarding the number of samples collected). One network that performed the text analysis had an input layer followed by seven fully connected layers (composed of 512, 256, 256, 256, 32, 10 and 4 nodes respectively). The second network had only one fully connected layer that was composed of four nodes. Following concatenation there was an additional fully connected layer with four nodes, prior to an output layer. Aside from the output layer, all layers employed rectified linear unit activation functions. The loss function was set as binary cross-entropy for classification experiments and mean absolute error for regression experiments.

Statistical analysis

Area under the receiver operator curve (AUC) was calculated using SciKit Learn. Other performance metrics, such as sensitivity, specificity and Youden's index, were calculated as standard [4, 5].

Statement of ethics

This project received approval from the relevant institutional ethics committee.

Results

Patient characteristics

There were 26,217 individual ward round notes included in this study, out of a maximum possible 26,322 (105 - 0.39% - excluded due to incomplete data). These ward round notes were from a total of 4,033 separate admissions, including a total of 3,412 unique patients. The mean LOS was 5.9 days (SD 7.8, minimum 0, maximum 136 days), median was 4 (IQR 2 – 7 days). 1,663 of the patients were female (48.7%) and the mean age was 67.8 years-old (SD 18.9). The number of ward round notes that were within two days of discharge was 7,432 (28.3%). The number of ward round notes that were within seven days of discharge was 17,224 (65.7%).

Repeated prediction of discharge in the next two days

In the prediction of discharge within the next two days, the artificial neural network was the best performing model (AUC 0.80) (Table 1). The logistic regression model achieved an

AUC of 0.78. When Youden's index was used as the cut-off score for the artificial neural network, sensitivity of 0.67, specificity of 0.79 and accuracy of 0.75 were achieved. The most common category in the test dataset was notes following which discharge did not occur within the next two days (64.9%). When a high sensitivity cut-off was employed, a high sensitivity (>0.95), was associated with a specificity of 0.27. With a high specificity cut-off (>0.95) sensitivity of 0.31 was returned.

Word stems associated with discharge within the next two days had good face validity. For example, word stems and n-grams that were most associated with discharge within the next two days included 'home today', 'discharg today', 'dc today', 'home tomorrow' and 'continu improv'. Similarly, word stems associated with not being discharged within the next two days had good face validity, with examples including 'dc plan', 'care await', 'sunday', 'sacat' (a widely used abbreviation for the South Australian Civil and Administrative Tribunal) and 'acut confus'.

Repeated prediction of discharge in the next seven days

The AUCs recorded in the prediction of discharge within the next seven days were lower. The artificial neural network achieved the highest AUC in prediction of discharge within the next seven days (0.68). Using Youden's index as the cut-off score, the artificial neural network provided a sensitivity of 0.64, specificity of 0.65 and accuracy of 0.64. The logistic regression model recorded an AUC 0.61. The most common category in the test dataset was notes following which discharge did occur within the following seven days (76.3%).

Repeated prediction of estimated date of discharge using ward round note

The artificial neural network achieved a mean absolute error of 3.9 days when predicting LOS from the time of the ward round note.

Discussion

This study has shown that, using mixed-data inputs available in daily ward round notes, ML can predict with reasonable performance which general medicine patients will leave within the next two days. Prediction of which patients will leave within seven days of the given note was somewhat effective. The generation of an accurate exact EDD remains inaccurate.

Although the medical text in ward round notes may include specific references to the timing of an expected discharge (for example, "Discharge home tomorrow"), it is not always the case that discharge will occur on that expected or planned day. Examples of possible causes for unexpected delays include clinical deterioration requiring ongoing admission, changes in patient preferences (for example regarding discharge destination) and unexpected delays with respect to logistics (such as transport or bed availability at discharge destinations). However, these limitations of medical text are not specific to text and would also be present should expected discharge date be recorded by doctors in a different fashion (such as a date-time field). The use of deep learning natural language processing may enable an automated hospital-wide process of estimating likelihood of discharge within the next two days (similarly to how a date-time field may enable the automated hospital-wide collection of such data), with the additional benefit of making data-driven adjustments to discharge likelihood based upon documented characteristics of an individual's given daily ward round review.

Given that medical text may only provide consistent insights into general medical LOS when discharge is likely to occur shortly after the documentation (e.g. within 48 hours), further inputs and strategies may be required to help predict LOS beyond this period. Additional inputs that could be sought include physiological parameters (such as vital signs, including when supplemental oxygen is not required), medications prescribed (such as when transitioned from intravenous to oral medications), pathology results (such as normalisation of a raised white cell count or creatine), radiology results (resolution of a pleural effusion or removal of nasogastric tubes for example) and nursing and allied health assessments (such as physiotherapy and occupational therapy notes). The incorporation of operation reports may also aid in LOS estimates, although the utility of such reports is likely to vary significantly between specialties and may be uncommon in general medicine.

This study is limited in that it was conducted at a single centre. It is conceivable that different patterns of medical text entry (for example different abbreviations) at different locations may limit the generalisability of the models. All text used in this study was in English. It should be noted that due to the nature of length of stay there was a degree of imbalance in the training and test datasets with respect to the proportion of cases in each category (see “Results” and Table 1). The performance of the models in this study should be viewed in the context of these imbalances in the test datasets.

In many instances utility of ML in medicine will be achieved through a synergistic approach with human decision making through the provision of additional information, rather than being a process conducted independently of human decision-making [6, 7]. The generation of EDDs may be one such area. Given the difficulty of producing accurate EDDs, studies may be considered in which the accuracy and timeliness of human-only EDD and human-with-ML

EDD may be compared. It is possible that ML may be able to provide a suggested timeframe for an EDD (e.g. between 2-7 days), and a clinician may then incorporate that information into their decision-making process when generating an EDD, thereby improving performance.

Future research in this area should aim to externally and prospectively validate models, such as those that performed well in this study, on datasets from other centres. Similar research studies in other specialty areas may also be conducted. Implementation studies are required that show improved patient or healthcare system outcomes, prior to the use of such models in routine clinical practice.

Conclusion

ML using mixed-data inputs, including daily ward round notes, appears effective in predicting when a patient will be discharged within 48 hours of the note being written. Prospective and external validation of these models are required. The generation of accurate EDDs remains inaccurate, and investigation of combination human-ML methods may be beneficial.

Conflict of interest**:** The authors have no conflicts of interest to declare.

Funding source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ou L, Chen J, Young L, Santiano N, Baramy L, Hillman K. Effective discharge planning – timely assignment of an estimated date of discharge. Australian Health Review. 2011;35:357-63.

[2] Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. Intern Emerg Med. 2020.

[3] Huang Z, Juarez JM, Duan H, Li H. Length of stay prediction for clinical treatment process using temporal similarity. Expert Systems with Applications. 2013;40:6330-9.

[4] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32-5.

[5] Hand DJ. Assessing the Performance of Classification Methods. International Statistical Review. 2012;80:400-14.

[6] Falavigna G, Costantino G, Furlan R, Quinn JV, Ungar A, Ippoliti R. Artificial neural networks and risk stratification in emergency departments. Internal and Emergency Medicine. 2018;14:291-9.

[7] Obermeyer Z, Emanuel E. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375:1216-9.

Figure 1: Schematic diagram of artificial neural network structure

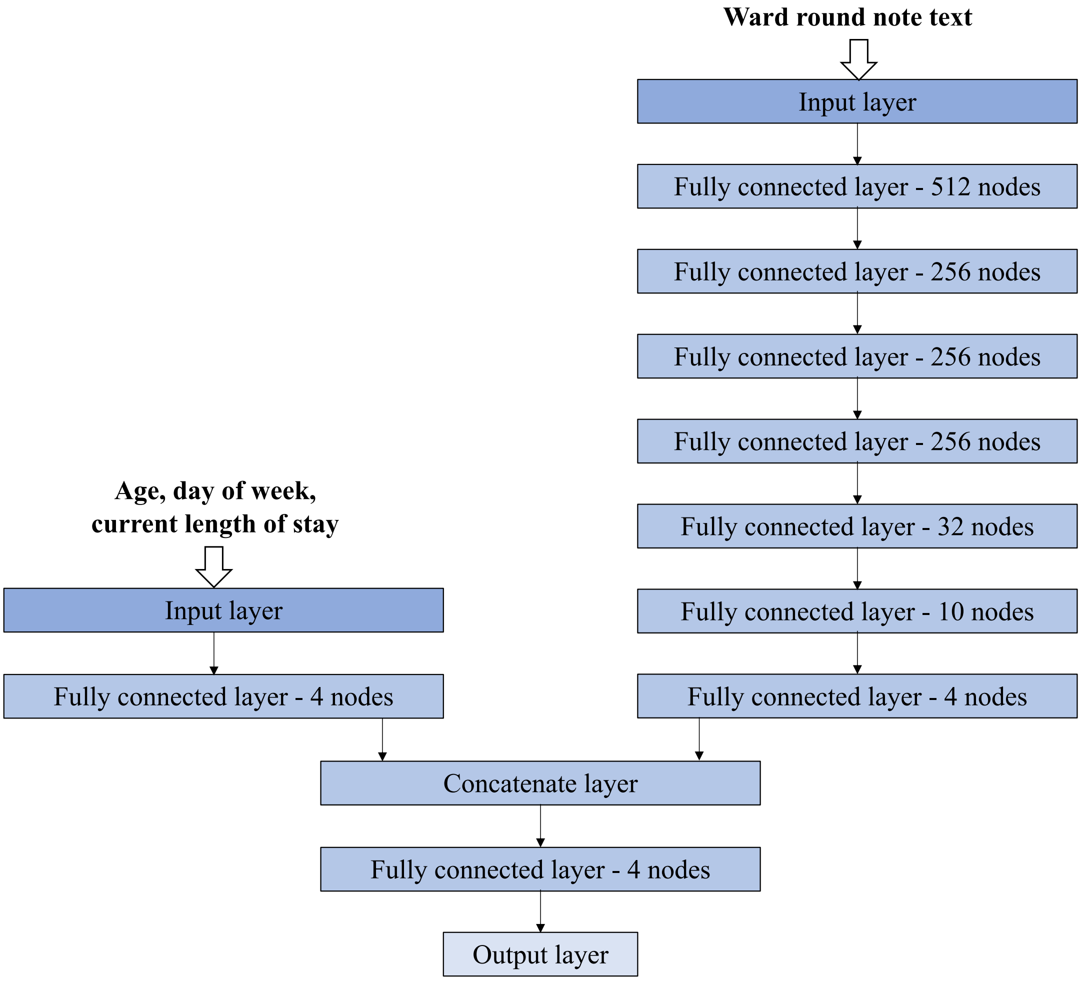

Ward round note text
Input layer
Fully connected layer - 512 nodes
Fully connected layer - 256 nodes
Fully connected layer - 256 nodes
Fully connected layer - 256 nodes
Fully connected layer - 32 nodes
Fully connected layer - 10 nodes
Fully connected layer - 4 nodes
Age, day of week,
current length of stay
Input layer
Fully connected layer - 4 nodes
Concatenate layer
Fully connected layer - 4 nodes
Output layer

Table 1: Results for machine learning models applied to prediction of discharge from time of ward round note

Model (Cut-off)	**AUC**	**PR score**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F1 Score**	**Accuracy**
Discharge in next 2 days												
ANN (Youden's index)	0.80	0.72	958	465	2081	556	0.67	0.79	0.63	0.82	0.65	0.75
ANN (High specificity)	0.80	0.72	447	976	2559	78	0.31	0.97	0.85	0.72	0.46	0.74
ANN (High sensitivity)	0.80	0.72	1361	62	707	1930	0.96	0.27	0.41	0.92	0.58	0.51
LR (Youden's index)	0.78	0.69	984	439	1907	730	0.69	0.72	0.57	0.81	0.63	0.71
Discharge in next 7 days												
ANN (Youden's index)	0.68	0.86	1971	1128	621	340	0.64	0.65	0.85	0.36	0.73	0.64
LR (Youden's index)	0.61	0.83	2002	1097	508	453	0.65	0.53	0.82	0.32	0.72	0.62

AUC = Area under the receiver-operator curve, PR= precision-recall, ANN = Artificial neural networks, LR= Logistic regression, LOS = Length of Stay, TP = True positive, FN = False negative, TN = True negative, FP = False positive, PPV = Positive predictive value, NPV = Negative predictive value.

Chapter 9 - Daily estimates of individual discharge likelihood with deep learning natural language processing in general medicine: a prospective and external validation study, *Internal and Emergency Medicine*

Citation

Bacchi S, Gilbert T, Gluck S, Cheng J, Tan Y, Chim I, Jannes J, Kleinig T & Koblar S 2021, 'Daily estimates of individual discharge likelihood with deep learning natural language processing in general medicine: a prospective and external validation study', *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02816-7

Statement of Authorship

Title of Paper	Daily estimates of individual discharge likelihood with deep learning natural language processing in general medicine: a prospective and external validation study
Publication status	▣ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication details	**Bacchi S**, Gilbert T, Gluck S, Cheng J, Tan Y, Chim I, Jannes J, Kleinig T & Koblar S 2021, 'Daily estimates of individual discharge likelihood with deep learning natural language processing in general medicine: a prospective and external validation study', *Internal Emerg Med*, https://doi.org/10.1007/s11739-021-02816-7

Principal Author

Name of Principal Author (Candidate)	Dr Stephen Bacchi		
Contribution to the Paper	Developed concept for project, designed methodology, gained relevant ethics and institutional approvals, performed data collection, performed data analysis, wrote report, submitted article and responded to reviewer comments.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	8/1/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Dr Toby Gilbert		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Sam Gluck		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Joy Cheng		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Yiran Tan		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Dr Ivana Chim		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	8/1/2022

Name of Co-Author	Prof Jim Jannes		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	31/1/2022

Name of Co-Author	Prof Timothy Kleinig		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	30/1/2022

Name of Co-Author	Prof Simon Koblar		
Contribution to the Paper	Critical appraisal of manuscript for important intellectual content.		
Signature		Date	9/3/2022

Abstract

Machine learning, in particular deep learning, may be able to assist with the prediction of the length of stay and timing of discharge for individual patients. Artificial neural networks applied to medical text have previously shown promise in this area. In this study, a previously derived artificial neural network was applied to prospective and external validation datasets. In the prediction of discharge within the next two days, when the algorithm was applied to prospective and external datasets, the area under the receiver operator curve for this task were 0.78 and 0.74 respectively. The performance in the prediction of discharge within the next seven days was more limited (area under the receiver operator curve 0.68 and 0.67). This study has shown that in prospective and external validation datasets the previously derived deep learning algorithms have demonstrated moderate performance in the prediction of which patients will be discharged within the next two days. Future studies may seek to further refine or evaluate the effect of the implementation of such algorithms.

Manuscript

Introduction

Discharge planning, including the prediction of length of stay to produce an estimated date of discharge, may reduce length of stay and readmission [8]. In medical inpatients it can be difficult to accurately predict length of stay [9]. It is possible that machine learning, in particular deep learning, may be able to assist with the prediction of patients' length of stay and timing of discharge.

The steps in the development of a deep learning algorithm for use in clinical medicine may be considered analogous to those involved in the development of a clinical decision rule or risk stratification score. These steps include pilot, derivation, validation, implementation and post-implementation studies [10, 11]. Previously, a pilot study demonstrated that medical officer documentation in the form of emergency department notes could feasibly be analysed by machine learning to provide an indicator of length of stay [2]. Subsequently, a derivation study was conducted in which an artificial neural network (ANN) demonstrated the best performance in the prediction of which patients would leave hospital within the next two days based upon the analysis of inputs, including the text of medical officer ward round notes [12].

The aims of this study were to evaluate the performance of the previously derived ANN when applied to (1) a prospectively collected validation dataset from the original center, and (2) an external validation dataset collected from another center.

Materials and Methods

Data collection

Data was collected from two hospitals for this study: location 1 (Royal Adelaide Hospital, Adelaide, Australia) for the prospective dataset, and location 2 (Queen Elizabeth Hospital, Adelaide, Australia) for the external validation dataset. Consecutive individuals admitted under General Medicine at location 1 in the four-month period following 1/9/2020 were included in the study. All individuals admitted to location 2 under General Medicine in the 12-month period following 1/1/2020 were included. The daily ward round notes of included individuals were collected for analysis, along with discrete variables including age, current length of stay, and the day of the week the ward round note was written.

Data pre-processing and deep learning models

Ward round notes with incomplete input or outcome data were excluded. Text data was pre-processed using the same method as that which was employed in the previous derivation study [12]. Briefly, negation detection was applied, punctuation removed, word stemming applied, stopwords removed, and n-grams of 1-3 word-stems in length were formed. Text was transformed with the count vectorizer that was fitted in the earlier study.

The artificial neural network structures, weights and cut-off scores developed in the derivation study were employed in this study [12]. In the previous study, 5-fold cross-validation was used to evaluate basic ANN and logistic regression models, and additional complexity (in the form of additional layers and nodes) was added progressively until model performance (as indicated by area under the receiver operator curve) no longer improved. The ANN was composed of two separate branches. The branch that received the discrete data had one input layer, followed by one fully connected layer with four nodes. The branch that received text data had an input layer followed by seven fully connected layers with 512, 256, 256, 256, 32, 10 and 4 nodes. These separate branches were merged and then followed by one fully connected layer and an output layer.

Classification experiments

The ANN and logistic regression models from the derivation study were applied to both the prospective dataset from location 1 and the external dataset from location 2. Youden's index, high specificity and high sensitivity cut-off scores were applied for predictions regarding discharge within the next two days. For predictions regarding discharge within the next seven days, only Youden's index was trialled.

Hospital-wide day-by-day prediction example

For the purposes of demonstration, high specificity and high sensitivity cut-off scores for the ANN, predicting discharge within the next two days, were applied on a day-by-day basis to all patients currently admitted on each given day at location 1. Using these two cut-off scores, a range for the predicted number of discharges within the next two days was generated. The number of individuals actually discharged within the next two-day period was calculated on a two-day rolling-basis.

Statistical analysis

The primary outcome was area under the receiver operator curve (AUC) for the prediction of discharge within the next two days on the prospective dataset. The AUC was calculated using the trapezoidal rule. Other performance characteristics including F1 score, sensitivity, specificity, positive predictive value and negative predictive value were also calculated.

Statement of ethics

Ethics approval was granted for this project by the Central Adelaide Local Health Network Research Ethics Committee.

Results

Patient characteristics

The dataset from location 1 included 16,550 individual notes, out of 16,606 identified notes (56 excluded due to incomplete data, 0.34%). These notes included information from 2,264 individuals and 2,612 separate hospital admissions. The average patient age was 71.4 (standard deviation 18.0). The median length of stay was 9 days (interquartile range 5 – 17 days) (see Table 1). 5,022 individual notes (30.3%) were within two days of discharge.

For location 2, 19,894 notes met inclusion criteria, of which 19,766 were used in the analysis (128, 0.64% excluded due to incomplete data). These notes were from 2,777 individuals over the course of 3,712 distinct admissions. The average age for this cohort was 74.8 (standard deviation 17.0). For this group the median length of stay was 8 days (interquartile range 4 – 15 days). 6,722 individual notes (34.0%) were within two days of discharge.

Prediction of discharge within two days

For location 1, an AUC of 0.78 was achieved in the prediction of discharge within the next two days (see Table 2). When the Youden’s index cut-off score was applied, this provided specificity for discharge within two days of 0.76, sensitivity of 0.67, positive predictive value of 0.55 and negative predictive value of 0.84. The high specificity cut-off score provided the highest accuracy of 0.75. The logistic regression model returned an AUC of 0.77.

When applied to the dataset from location 2, an AUC of 0.74 was achieved. With the application of the Youden's index cut-off score, specificity of 0.77, sensitivity 0.60, positive predictive value 0.57 and negative predictive value 0.79 were returned. The high sensitivity cut-off score provided sensitivity 0.95, and specificity 0.16. The high specificity cut-off score provided specificity 0.98, and sensitivity 0.18. For location 2, the logistic regression model provided an AUC of 0.73.

For aspects of analysis regarding interpretability, please refer to the previous study [12]. Briefly, word stems and n-grams most associated with a high (or low) likelihood of discharge in the next two days had good face validity. For example, references to certain symptoms (such as acute confusion) or particular aspects of discharge planning (such as tribunals) were associated with a lower likelihood of discharge in the next two days.

Prediction of discharge within seven days

The application of the ANN to the prediction of discharge within seven days of the given ward round note, returned an AUC of 0.68 for location 1 and 0.67 for location 2. With the application of Youden's index cut-off score, accuracies of 0.56 and 0.54 were returned for location 1 and 2 respectively (noting that the proportion of individuals in the most common class in the dataset comprised 69% and 72% for this outcome in location 1 and 2). The logistic regression model provided AUC of 0.68 and 0.66 for locations 1 and 2 respectively.

Hospital-wide day-by-day prediction example

The actual number of discharges on a hospital-wide basis fell within the predicted range (derived from high sensitivity and high specificity cut-off scores applied to individual ward round notes) for 107/121 (88.4%) of the two-day periods included for location 1 (see Figure

1). Note that for each time-point (X axis) in Figure 1, there is an actual number of discharges and range for the number of discharges that was predicted (Y axis), and Figure 1 starts with a low number of actual and predicted discharges in the first instance. This is because at the commencement of the inclusion period, the majority of inpatients currently admitted had been admitted prior to the time at which they would meet the criteria for inclusion.

Discussion

This study has demonstrated stable performance of an ANN for the prediction of discharge within the next two days on prospective and external validation datasets (0.78 and 0.74). The AUC achieved on the test dataset in the derivation study was 0.80 [12]. The prediction of discharge within seven days also had stable, although limited, performance.

Variable cut-off scores and aspects of patient population characteristics may influence model performance. It is seen as one of the potential benefits of models such as those in this study that there is the ability to develop and employ variable cut-off scores, which may improve potential usefulness for some applications (as demonstrated in the hospital-wide day-by-day prediction example). As for any other test, in a population with a different prevalence of the condition of interest (in this case patients who are discharged within the next two days), the model would have correspondingly different positive and negative predictive value as outlined by Bayes' theorem [13].

The potential use cases for algorithms such as those employed in this study may be either at the level of the individual or the healthcare system. At an individual level, potential utility could include automatically highlighting those with a high probability of impending discharge to staff involved in organising discharge (such as ward pharmacists and staff organising transport), and prompting closer examination of barriers to discharge in individuals who are planned for discharge, but are predicted to have a low probability of leaving. At a system level possible uses could include helping to automatically identify ward beds which may soon become available to assist with bed flow, and pre-emptively identifying periods during which hospital bed occupancy may be high due to a lack of discharges. At this stage, such possible applications are hypothetical, and improved outcomes need to be demonstrated in implementation studies.

An additional consideration regarding the future utility of such models is the incorporation of human length of stay estimates. It may be hypothesised that human length of stay estimates in which the doctor was informed of the model's prediction, or models that incorporate human length of stay estimates, may have better performance than either human or machine learning estimates alone. Such an approach would support the use of machine learning as an adjunct to improve, rather than replace, human decision making, as has been discussed previously in other areas of medicine [6]. In this study, the performance of the models was not able to be compared to human estimates. Future comparisons to and the incorporation of human length of stay estimates are warranted.

Although the strengths of this study include the use of both prospective and external validation datasets, there are several limitations that should be acknowledged. For example, all of the text analysed in this study was in English. Both sites involved in the study are based within the same Local Healthcare Network, with a significant overlap in the junior and senior medical staff who are responsible for authoring the notes. Centre-wise comparisons of factors that may influence length of stay (such as availability of discharge facilities) were beyond the scope of the project. There is a degree of imbalance in the validation dataset classes included in this study. For location 1 and location 2 respectively, 30.3% and 34.0% of notes had discharge occur within the next two days. This type of imbalance may be inherent when using length of stay as an outcome but should still be taken into account when evaluating model performance (see Table 2).

Future strategies to improve the performance and potential utility of models could include the investigation of additional ANN topologies, such as recurrent neural networks and radial basis function networks [14]. For example, recurrent neural networks have demonstrated robust performance in the prediction of the development of in-hospital acute kidney injury [15]. The use of additional topologies could also facilitate and be investigated in the prediction of other important outcomes, such as hospital readmission [16].

Further research is warranted. Future studies may seek to develop similar models for medical documentation in languages other than English, or in other subspecialties. Using medical notes from the days prior to the date on which the prediction is made, including notes from nursing and allied health professionals and the addition of pathology, radiology and pharmacy data, may improve the accuracy of these models. Implementation studies are required to demonstrate improvement in patient or system outcomes prior to routine use of such algorithms.

Conclusions

In prospective and external validation datasets, previously derived deep learning algorithms analysing text from medical notes have demonstrated stable performance in the prediction of which patients will be discharged within the next two days. Future studies may seek to evaluate the effect of the implementation of such algorithms.

Conflict of interest**:** The authors have no conflicts of interest to declare.

Funding source: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ou L, Chen J, Young L, Santiano N, Baramy L, Hillman K. Effective discharge planning – timely assignment of an estimated date of discharge. Australian Health Review. 2011;35:357-63.

[2] Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. Intern Emerg Med. 2020.

[3] Huang Z, Juarez JM, Duan H, Li H. Length of stay prediction for clinical treatment process using temporal similarity. Expert Systems with Applications. 2013;40:6330-9.

[4] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32-5.

[5] Hand DJ. Assessing the Performance of Classification Methods. International Statistical Review. 2012;80:400-14.

[6] Falavigna G, Costantino G, Furlan R, Quinn JV, Ungar A, Ippoliti R. Artificial neural networks and risk stratification in emergency departments. Internal and Emergency Medicine. 2018;14:291-9.

[7] Obermeyer Z, Emanuel E. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016;375:1216-9.

[8] Goncalves-Bradley DC, Lannin NA, Clemson LM, Cameron ID, Shepperd S. Discharge planning from hospital. Cochrane Database Syst Rev. 2016:CD000313.

[9] Bacchi S, Tan Y, Oakden-Rayner L, Jannes J, Kleinig T, Koblar S. Machine Learning in the Prediction of Medical Inpatient Length of Stay. Internal Medicine Journal. 2020.

[10] Stiell IG, Wells GA. Methodologic Standards for the Development of Clinical Decision Rules in Emergency Medicine. Ann Emerg Med. 1999;33:437-47.

[11] Stiell IG, Bennett C. Implementation of Clinical Decision Rules in the Emergency Department. Academic Emergency Medicine. 2007;14:955-9.

[12] Bacchi S, Gluck S, Tan Y, Chim I, Cheng J, Gilbert T, et al. Mixed-data deep learning in repeated predictions of general medicine length of stay: a derivation study. Intern Emerg Med. 2021;16:1613-17.

[13] Grunau G, Linn S. Commentary: Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. Front Public Health. 2018;6:256.

[14] Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PLoS One. 2019;14:e0212356.

[15] Kim K, Yang H, Yi J, Son HE, Ryu JY, Kim YC, et al. Real-Time Clinical Decision Support Based on Recurrent Neural Networks for In-Hospital Acute Kidney Injury: External Validation and Model Interpretation. J Med Internet Res. 2021;23:e24120.

[16] Huang Y, Talwar A, Chatterjee S, Aparasu RR. Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC Med Res Methodol. 2021;21:96.

Figure 1: Demonstration of the use of high sensitivity and high specificity cut-off scores to generate a day-by-day predicted range for the absolute number of discharges in the next two days at a single centre over the study period.

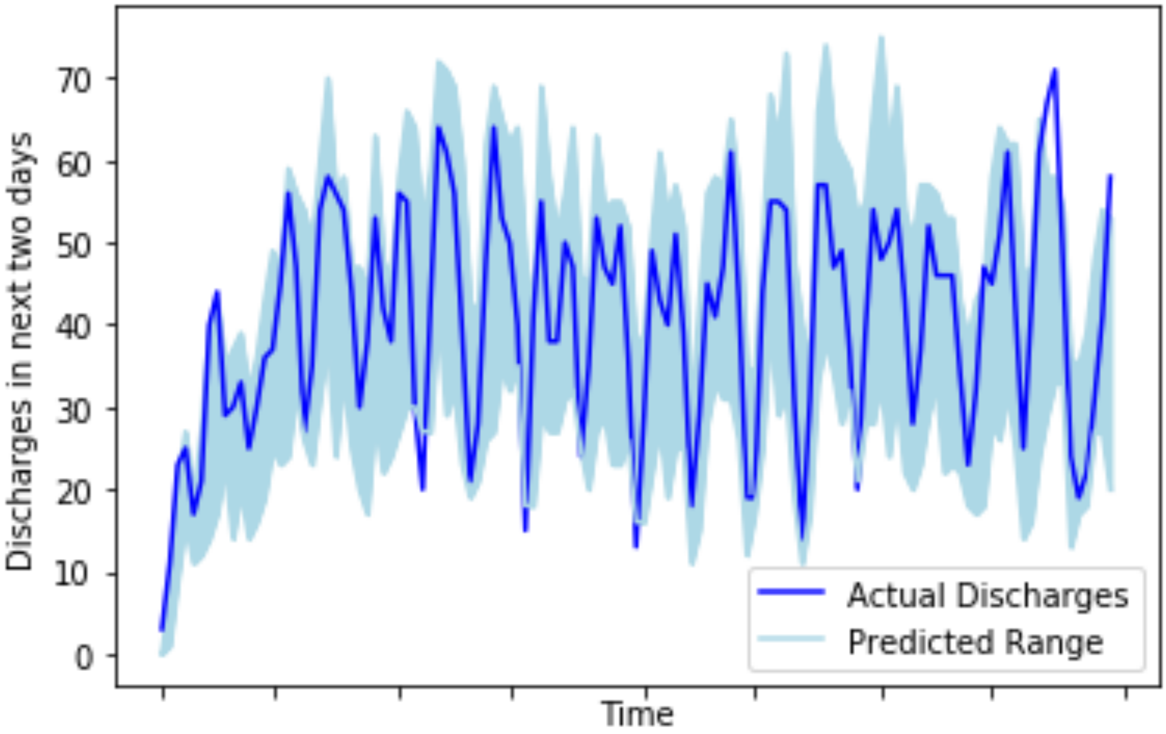


Table 1: Patient characteristics in prospective and external validation datasets

Characteristics	**Location 1 (prospective validation)**	**Location 2 (external validation)**
Number of included notes	16,550	19,766
Number of distinct admissions	2,612	3,712
Mean age (standard deviation)	71.4 years (18.0)	74.8 years (17.0)
Median length of stay (interquartile range)	9 days (5 - 17)	8 days (4 – 15)

Table 2**:** Performance of artificial neural networks in the prediction of length of stay in prospective and external datasets

Outcome	**Model**	**Cut-off**	**AUC**	**PR score**	**TP**	**FN**	**TN**	**FP**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F1 Score**	**Accuracy**
Location 1 (prospective validation)														
Discharge next 2 days	Artificial neural network	High specificity	0.78	0.65	1093	3929	11302	226	0.22	0.98	0.83	0.74	0.34	0.75
		Youden's index			3384	1638	8728	2800	0.67	0.76	0.55	0.84	0.60	0.73
		High sensitivity			4888	134	1596	9932	0.97	0.14	0.33	0.92	0.49	0.39
	Logistic regression	Youden’s index	0.77	0.63	3309	1713	8538	2990	0.66	0.74	0.53	0.83	0.59	0.72
Discharge next 7 days	Artificial neural network	Youden's index	0.68	0.84	5216	6234	4132	968	0.46	0.81	0.84	0.40	0.59	0.56
	Logistic regression	Youden’s index	0.68	0.82	7868	3582	2844	2256	0.69	0.56	0.78	0.44	0.73	0.64
Location 2 (external validation)														

Discharge next 2 days	Artificial neural network	High specificity	0.74	0.63	1178	5544	12799	245	0.18	0.98	0.83	0.70	0.29	0.71
		Youden's index			4011	2711	10015	3029	0.60	0.77	0.57	0.79	0.58	0.71
		High sensitivity			6396	326	2074	10970	0.95	0.16	0.37	0.86	0.53	0.43
	Logistic regression	Youden’s index	0.73	0.62	3966	2756	9945	3099	0.59	0.76	0.56	0.78	0.58	0.70
Discharge next 7 days	Artificial neural network	Youden's index	0.67	0.84	6049	8162	4564	991	0.43	0.82	0.86	0.36	0.57	0.54
	Logistic regression	Youden’s index	0.66	0.83	9008	5203	3314	2241	0.63	0.60	0.80	0.39	0.71	0.62

Chapter 10 - General Discussion and Future Perspectives

This thesis has described the derivation and validation of machine learning models, in particular deep learning models, for the prediction of clinically significant outcomes in stroke and general medicine. The outcomes predicted include inpatient outcomes related to discharge planning, key performance indicator (KPI) data extraction and clinical coding categorisations. This research has contributed significantly to the existing literature for stroke and general medicine predictive analytics through the application of methods such as open-source natural language processing and recurrent prediction generation. Strategies to improve the utility of these models may include leveraging variable cut-off scores, integration into electronic medical records, and combining model predictions with human predictions. Future research in this area seeking to implement these models to examine their effect on patient and system-oriented outcomes are warranted.

Using predictive analytics for inpatient outcomes for discharge planning, KPIs and clinical coding is a significant contribution to the existing stroke literature

Machine learning and related predictive analytics strategies have been employed previously in stroke [1]. A significant focus of these previous studies has been on prognostication through analysis of acute stroke imaging [2, 3]. However, this type of imaging analysis alone is unable to capture all aspects of a patient's presentation that guides their management decisions [4]. In particular, analysis of imaging alone would not necessarily be able to capture an individual's baseline functional status, exact neurological deficit and post-stroke functional deficit (e.g., with respect to swallowing and mobility, which may be affected by comorbidities other than the stroke itself). Accordingly, predictive analytics methods for prognostication using clinical, demographic and comorbidity information are required [5].

Previous prognostic scores using such characteristics have focussed on the prediction of outpatient outcomes (such as 3-month mortality) rather than inpatient outcomes (such as length of stay, independence at discharge and discharge destination) [6]. Chapter 2 of the thesis is a literature review examining the previous applications of machine learning in the prediction of medical (including stroke) inpatient length of stay. This review demonstrates that machine learning has rarely been applied to length of stay prediction in stroke (and general medicine), and that the existing studies typically lack prospective and external validation studies.

Chapter 3 describes a derivation study for the development of machine learning models in the prediction of stroke inpatient outcomes to assist with discharge planning, using only data available at the time of admission. In this study of 2,840 patients the prediction of functional independence at the time of discharge, survival to discharge and discharge to home was successful using logistic regression models and artificial neural networks. The area under the receiver operator curve (AUC) for these outcomes were 0.90, 0.90 and 0.81 respectively. The prediction of length of stay ≤8 days had lower performance with an AUC of 0.67. The most likely explanation as to why length of stay was unable to be more effectively predicted is that factors significantly affecting length of stay were not represented in the input data. Such factors may include the insurance status of the patients, the bed-state at the time of admission (i.e. system overcrowding and access block), and progress throughout the hospital stay after the time of admission. The results for the prediction of these outcomes were similar in the prospective and external validation study described in Chapter 4. Namely, classification performance was sound in the prediction of discharge independence (AUC 0.85 and 0.87), survival to discharge (AUC 0.91 and 0.92), and discharge destination (AUC 0.76 and 0.78) for both the prospective and external validation datasets respectively. The performance in prediction of length of stay ≤8 days remained limited (AUC 0.62 and 0.66).

Collection of KPI data and correct classification with clinical coding are important for stroke medicine in multiple forms, including quality assurance, as well as guiding service provision planning and activity-based funding [7, 8]. In Chapter 5 a variety of techniques are employed to demonstrate that automated data extraction from medical free-text can be performed with predictive analytics. In particular, random forest models achieved a high level of classification performance in classification of type of stroke (AUC 1.00), whether an individual received thrombolysis on-site (AUC 0.97), whether an individual received endovascular thrombectomy (AUC 1.00), and whether an individual had atrial fibrillation/flutter (AUC 0.97). Natural language processing methods have been employed previously in the field of clinical coding [9]. However, the majority of this software is proprietary in nature. Chapter 6 describes the application of open-source software to the task of stroke clinical coding. The use of this open-source software was able to correctly identify instances in which the original clinical coding had misclassified Adjacent Diagnosis Related Groups, International Statistical Classification of Diseases and Related Health Problems classifications, and Diagnosis Related Groups.

When a classifier achieves perfect classification performance (AUC 1.00 or accuracy 1.00) this result raises the possibility of overfitting. For example, the possibility of overfitting may be considered for the random forest in the detection of endovascular thrombectomy in Chapter 5. However, it should be noted that in Chapter 6 this same model was applied to a non-overlapping dataset and achieved a classification accuracy of 0.974. The generalisability of the model to a separate dataset can be seen to make the possibility of overfitting less likely.

In these studies, the models employed included logistic regression, decision tree, random forest and artificial neural network algorithms. These models were selected for a number of reasons. In particular, all outcomes for prediction in these studies were categorial and binary in nature. This aspect of the analysis lends itself to the use of logistic regression. There were instances of unbalanced datasets, particularly with respect to KPI data. Decision tree algorithms and random forest algorithms are known to perform comparatively well on unbalanced datasets. Deep learning has provided advances in machine learning classification performance in medicine, particularly in the analysis of large datasets. The presence of large datasets prompted the use of artificial neural networks in this context. In the initial stroke inpatient outcome prediction study (Chapter 3), all four types of models were evaluated. The best performing models, logistic regression and artificial neural networks, were then used again in the subsequent validation study (Chapter 4). Similarly, all four models were employed in the analysis of stroke KPIs (Chapter 5), with the best performing model, the random forest model, subsequently applied again in the analysis of stroke clinical coding data (Chapter 6). These models were developed using open-source Python libraries, namely TensorFlow, NLTK and SciKit-Learn [10-12]. These libraries were selected due to their open-source nature and widespread use in the machine learning community. The rationale for this choice is that, although the prospective and external validation studies were successful, for other centres the use of local data may enable the development of centre-specific algorithms. Accordingly, the use of open-source and widely available libraries would facilitate the development of centre-specific algorithms at other locations.

Recurrent prediction with mixed-data inputs is a novel method for predicting outcomes in the complex general medicine patient population

General medicine comprises a wide array of patients with complex comorbidities and social issues [13]. Discharge planning is important for this group. Although attempts have been made to develop tools, few studies have applied machine learning, and in particular deep learning, specifically to this group [14]. Machine learning research that is applicable to the group has included all hospital inpatients; however, such studies are typically focussed on "early warning systems" that predict inpatient mortality or intensive care unit admission [15]. Chapter 2 outlines a literature review that examines the previous use of machine learning methodologies to predict medical inpatient length of stay. No identified studies specifically examined the general medicine population.

In Chapters 7, 8 and 9 studies are conducted that pilot, derive and then validate machine learning models for the prediction of timing of discharge for general medicine inpatients, to assist with discharge planning. In the pilot study in Chapter 7 natural language processing, namely an artificial neural network, is applied to emergency department notes, and it is found that the models had moderate performance in the prediction of which patients would be discharged in ≤2 days or >2 days (AUC 0.75). In the prediction of an actual date of discharge, model performance was poor. Chapter 8 describes a derivation study that employs machine learning methods using both natural language processing and discrete data fields, to make daily predictions regarding an individual's likelihood of discharge in the next 48 hours. This approach was successful, with the artificial neural network achieving an AUC of 0.80. The prediction of discharge within 7 days was less accurate (AUC 0.68). Finally, in Chapter 9, these models underwent prospective and external validation. In this prospective and external validation study, the models again demonstrated potentially useful performance (AUC 0.78 and 0.74) in the prediction of discharge within the next 48 hours.

In these studies, it was found that certain words and word stems that explicitly made reference to discharge plans (such as “Discharge home tomorrow”) were often predictive of discharge within the next 48 hours. However, it should be noted that in clinical practice discharges do not always occur as planned. For hospital inpatients there are many possible reasons for unexpected delays of discharge including, but not limited to, clinical deterioration (for example, newly hospital-acquired complications that may necessitate ongoing admission), changes in patient preferences (for example, with respect to preferred discharge date or destination) and logistical issues (for example, delays in diagnostic testing, treatments, transport or changes in discharge destination bed availability). It is in the context of this unpredictability that the performance of the algorithms should be viewed.

The level of classification performance on a given task that is considered highly accurate and/or useful depends on the task in question. In tasks that are straightforward, a significantly higher level of classification performance would be expected than in complex and nuanced tasks such as predicting the timing of discharge. It is possible that there will never be a means of predicting patient discharge 48 hours in advance with complete accuracy. Similarly, an acceptable or useful level of performance is dependent on the nature of the task. This threshold may depend on multiple factors, including the difficulty of predicting a given outcome and the consequences of misclassification.

The models developed in the initial study in this area (Chapter 7) included logistic regression, random forest, artificial neural network and convolutional neural networks. These models were selected on the basis of the results from previous analyses of medical free-text, in addition to the rationale outlined for their use in Chapters 3 and 5, as discussed above [16]. The best performing types of models from this initial study, were then used again in Chapters 8 and 9. These studies also employed open-source Python libraries for machine learning analyses, namely TensorFlow, NLTK and SciKit-Learn [10-12]. The reason for this selection was to improve generalisability and facilitate future analogous studies by other researchers. These widely available resources could be used by others in combination with local data to develop centre-specific algorithms. Therefore, the use of open-source and widely available libraries may facilitate this future development of algorithms at other centres.

Strategies to facilitate the potential utility of these models may include the use of variable cut-off scores, integration into electronic medical records and combination with human estimates

There are a variety of strategies that may be employed to improve the potential utility of the models developed and validated in the research in this thesis. These strategies may include the use of variable cut-off scores, integration with electronic medical records, and combining machine learning and human predictions. Additionally, diverse centres may benefit from deriving similar models locally by applying the methods in this research to local data.

Using variable cut-off scores, the classification models can produce predictions with varying levels of sensitivity and specificity. The use of these variable cut-off scores is possible because the output of the classification models in this thesis, prior to the application of a cut-off score, may be viewed as the probability of a positive result [17, 18]. Selecting different cut-off scores can be used to alter the sensitivity and specificity of the model outputs [19]. For example, to provide a prediction with high specificity, a high cut-off score could be selected. This cut-off score would provide greater specificity at the cost of reduced sensitivity. By using multiple cut-off scores for the output of an individual model a range of predicted outcomes can be produced, as is demonstrated in the hospital-wide day-by-day example in Chapter 9.

Additionally, rather than employing one or more cut-off scores, the model probability output could be used to rank individuals based on the likelihood of a positive result. For example, in the identification of potential misclassifications of clinical coding Diagnosis Related Groups (as in Chapter 6), instances of potential misclassification can be ranked based on the likelihood of a misclassification being present (rather than simply providing a binary outcome suggesting a misclassification is present or absent). Similarly, currently admitted patients could be ranked by their estimated probability of discharge within the next 48 hours (for example, if trying to identify those who may be most amenable to an intervention to facilitate earlier discharge). By employing this strategy an individual reviewing the outcome of the model may start with the instances in which the probability of the target result is highest. This technique may be useful when time or resources are limited.

Integration with electronic medical records would improve the potential utility of the developed models. As discussed in Chapter 1, prognostic scores that require the entry of many discrete data fields may be cumbersome and limit adoption and utility. An advantage of the utilised natural language processing methodologies is that data can be collected and analysed on potentially all comorbidities from a single input: medical free-text. Even so, when predictive analytics are to be employed at a large scale, and in high-demand resource-limited settings, streamlining the process by which data is entered and predictions presented would be beneficial. The ideal way to streamline this process would be through integration with electronic medical records, which is an area of ongoing development [20, 21]. Depending on the developer software and user access for given electronic medical records, it would be feasible to automatically generate daily predictions of the likelihood of discharge in the next 48 hours and present these predictions, either individually or as compiled reports, without any human time investment after installation. Similarly, integration into the electronic medical record would facilitate the automatic collection of KPI data. In order for KPI data to be acted upon and influence patient care it is necessary for the data to be available in near-real-time. Automatic data collection for research purposes could likely be asynchronous in the majority of cases. However, electronic medical record integration for this purpose is still of significant importance to reduce the number of intermediary steps involved in the data collection process.

It has been demonstrated that unnecessary alerts and time-consuming aspects of electronic medical record use may contribute to alert fatigue and burnout [22, 23]. These issues should be taken into consideration when considering how to optimise the potential utility of the models. The means by which healthcare professionals are notified of machine learning predictions should be constructed to minimise unnecessary alerts or burdensome mandatory electronic medical record fields.

By combining machine learning and human predictions it is possible that performance may be greater than either approach in isolation. While many studies compare machine learning performance to human performance, such comparisons may have limited usefulness, particularly with respect to factors relevant to discharge planning. As in other fields, it is considered unlikely that machine learning will guide decisions regarding discharge independent of human decision-making in the near future and, instead, it is more likely that machine learning may be employed *in conjunction* with clinician decision-making to help guide these processes [24]. It is possible that by implementing a hybrid approach, utilising machine learning-informed clinician decision-making, the utility of the models will be further improved. As described in Chapter 8, one possible example of how machine learning may be employed in conjunction with clinician decision-making would be for the model to suggest a timeframe for an estimated discharge date (e.g., between 2-7 days), and a clinician then be provided with the estimated range prior to selecting their chosen estimated discharge date. Such a synergistic approach may be hypothesised to improve the accuracy of estimated discharge date generation, completeness and timeliness.

Although the prospective and external validation studies were successful, it is possible that machine learning models derived in this thesis may not generalise to some diverse centres. In the event that the models did not generalise to a diverse centre, the application of the methods in this research to local data may enable the development of centre-specific algorithms. For example, the same stroke admission variables could be collected for admissions at the diverse centre, and logistic regression models then be derived from local data to produce centre-specific algorithms. Similarly, if a centre had a non-English primary written language, and translation strategies were unsuccessful, the same pre-processing and neural network structure could be applied to local non-English ward round notes to derive local models. In addition to employing the methods in this research, it is possible that the use of the algorithms from this research could facilitate the use of transfer learning (the process by which a pre-trained model from a different dataset can be retrained and applied to a new dataset, subsequently improving performance) to develop centre-specific algorithms at diverse centres.

Limitations of this research should be acknowledged

The limitations of each individual study are discussed in its corresponding chapter. Common to all natural language processing studies was the fact that in this research all analyses were conducted on English text (Chapters 5, 6, 7, 8 and 9). In addition, external validation studies (Chapter 4 and Chapter 9), while at distinct external centres, were all conducted within South Australia. Accordingly, application of the model to datasets from more diverse centres may be beneficial.

It should be noted that certain datasets in these studies were unbalanced. This issue is common in medical machine learning. KPI data may become particularly unbalanced when adherence to KPIs is strict. Strategies used to mitigate this issue included large datasets, the presentation of multiple performance metrics (rather than only accuracy) and the use of decision tree-based algorithms. If more unbalanced datasets are encountered in future analyses, for example for KPIs with adherence nearing 100%, additional methods that may be employed include resampling strategies and the generation of synthetic data.

It has been demonstrated in previous studies and reviews that machine learning has the potential to perpetuate existing biases. There are multiple means by which algorithms can become biased, including composition of training datasets, data annotation, as well as the applications to which the models are applied [25]. In particular, imbalances in datasets with respect to country of origin or demographics, such as gender, may bias machine learning models [26, 27]. Strategies exist to reduce this potential for bias such as the use of diverse and representative datasets [28, 29]. The studies in this thesis included large representative datasets from the local population, which reduce the potential for bias in this setting. However, applied in other diverse settings for which the dataset is less representative, there would be more potential for bias. As a component of implementation studies, ongoing post-implementation monitoring for potential biases is important. Application of algorithms to other datasets may also highlight potential biases.

Future research implementing models developed in these and similar studies is warranted

Future research in this area may involve implementation studies employing the existing models, and further studies aiming to improve the performance of the previously derived models. Implementation studies should aim to examine patient and system-oriented outcomes [30]. These studies should also adhere to guidelines regarding artificial intelligence trials [31]. Approaches that may be examined to improve existing model performance include the use of novel neural network topologies and the use of additional input data (both individual data and system-based data).

Implementation studies are required to demonstrate improvement in meaningful clinical outcomes with model deployment. Meaningful outcomes may include benefits in patient flow, workforce usage and cost-benefit analyses. There are now published guidelines for the development of artificial intelligence clinical trials [31, 32]. However, the design of these implementation studies will require careful consideration. Implementation studies using machine learning to aid with discharge planning will likely need to compare clinician *with* machine learning interventions against a comparator of clinicians *without* machine learning. Accordingly, trial design considerations with respect to possible blinding of clinicians may be difficult [33]. This difficulty with blinding highlights the importance of other trial design elements aimed to reduce bias (such as randomisation, as opposed to pseudo-randomisation). It is fundamental to such clinical trials that the pre-specified endpoints are clinically relevant.

Such clinically relevant endpoints may be significant at an individual level, or at a system-level (such as reduction in length of stay). If the model deployment improves outcomes in stroke and general medicine patients, future research may seek to employ similar methods in other specialties.

Implementation study outcomes will necessarily take into consideration the logistical components of hospital medical practice. Outcomes including total length of stay are important. However, additional outcomes, such as the proportion of discharge scripts and transport organised the day prior to discharge, may also be evaluated. These outcomes would be particularly meaningful if they alter the patients' experience of the discharge process in a positive manner. Patient-reported outcomes may also be an important endpoint to evaluate in implementation studies. The evaluation of KPI adherence may be targeted as a potential outcome of implementation studies, particularly if the collection of KPI data can be automated and notifications sent if it appears likely that KPIs will not be met.

Implementation studies in this area will need to consider the entire allied health team. Discharge planning is a multidisciplinary process. Machine learning predictions based on the ward round notes written by doctors may inform the practice of multiple team members, including nurses, pharmacists, and physiotherapists. Future studies may also apply machine learning to the notes written by these healthcare professionals to evaluate how the analysis of these notes may influence aspects of outcome prediction and discharge planning.

The use of synthetic data may be one avenue to improve the performance of the models developed in these studies. The use of synthetic data could increase the sample size for rare and uncommon medical conditions, and therefore reduce the likelihood of misclassification due to underrepresentation in the training datasets [34]. Synthetic data may be generated through multiple methods including data perturbation and generative adversarial networks [35].

The development and refinement of medical machine learning models, such as those described in this thesis, will be an ongoing iterative process. For example, future models may seek to utilise additional input data. This data may take the form of more data of the existing types used in this thesis (for example larger and more diverse datasets of ward round notes), as well as new types of data. New data that may be incorporated into models include individual data (such as laboratory test results and imaging results) as well as system data (such as the availability of beds at nearby rehabilitation centres). However, it should be noted that individual data (such as salient laboratory test results) are often typed or copied into ward round notes, and may therefore not necessarily add significantly to model performance. The use of additional model structures may also be investigated in this area in future. Over time, novel machine learning methods (such as additional artificial neural network topologies including recurrent and recursive neural networks) will continue to be developed [36]. Accompanying these novel machine learning methods, further pre-trained models (such as Bidirectional Encoder Representations from Transformers) will also become available that may be employed with transfer learning [37, 38]. Accordingly, further research in this area may seek to employ new data and methods, as they become available, to build upon the already successful models derived and validated in this thesis.

References

[1] Sirsat MS, Ferme E, Camara J. Machine Learning for Brain Stroke: A Review. J Stroke Cerebrovasc Dis. 2020;29:105162.

[2] Bivard A, Churilov L, Parsons M. Artificial intelligence for decision support in acute stroke - current roles and potential. Nat Rev Neurol. 2020;16:575-85.

[3] Kamal H, Lopez V, Sheth SA. Machine Learning in Acute Ischemic Stroke Neuroimaging. Front Neurol. 2018;9:945.

[4] Liberman AL, Pinto D, Rostanski SK, Labovitz DL, Naidech AM, Prabhakaran S. Clinical Decision-Making for Thrombolysis of Acute Minor Stroke Using Adaptive Conjoint Analysis. Neurohospitalist. 2019;9:9-14.

[5] Saposnik G, Johnston SC. Decision making in acute stroke care: learning from neuroeconomics, neuromarketing, and poker players. Stroke. 2014;45:2144-50.

[6] Drozdowska BA, Singh S, Quinn TJ. Thinking About the Future: A Review of Prognostic Scales Used in Acute Stroke. Front Neurol. 2019;10:274.

[7] Urimubenshi G, Langhorne P, Cadilhac DA, Kagwiza JN, Wu O. Association between patient outcomes and key performance indicators of stroke care quality: A systematic review and meta-analysis. Eur Stroke J. 2017;2:287-307.

[8] Independent Hospital Pricing Authority. Australian Refined Diagnosis Related Groups (AR-DRGs). 2020.

[9] Campbell S, Giadresco K. Computer-assisted clinical coding: A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. Health Inf Manag. 2020;49:5-18.

[10] Bird S, Klein E, Loper E. Natural Language Processing with Python: O’Reilly Media Inc.; 2009.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825−30.

[12] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16). 2016.

[13] Nardi R, Scanelli G, Corrao S, Iori I, Mathieu G, Cataldi Amatrian R. Co-morbidity does not reflect complexity in internal medicine patients. Eur J Intern Med. 2007;18:359-68.

[14] D'Souza AN, Said CM, Leggett NE, Tomkins MS, Kay JE, Granger CL. Assessment tools and factors used to predict discharge from acute general medical wards: a systematic review. Disabil Rehabil. 2021:1-15.

[15] Nannan Panday RS, Minderhoud TC, Alam N, Nanayakkara PWB. Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): A narrative review. Eur J Intern Med. 2017;45:20-31.

[16] Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep Learning Natural Language Processing Successfully Predicts the Cerebrovascular Cause of Transient Ischemic Attack-Like Presentations. Stroke. 2019;50:758-60.

[17] Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics. 2018;16:17-32.

[18] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020;9:14.

[19] Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochem Med (Zagreb). 2016;26:297-307.

[20] Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25:1419-28.

[21] Davenport T, Hongsermeier T, McCord K. Using AI to Improve Electronic Health Records. Harvard Business Review; 2018.

[22] Rush JL, Ibrahim J, Saul K, Brodell RT. Improving Patient Safety by Combating Alert Fatigue. J Grad Med Educ. 2016;8:620-1.

[23] Eschenroeder HC, Manzione LC, Adler-Milstein J, Bice C, Cash R, Duda C, et al. Associations of physician burnout with organizational electronic health record support and after-hours charting. J Am Med Inform Assoc. 2021;28:960-6.

[24] Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702.

[25] Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Language and Linguistics Compass. 2021;15.

[26] Fraser HS, Celi LA, Cellini J, Charpignon M-L, Dee EC, Dernoncourt F, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. PLOS Digital Health. 2022;1.

[27] Agmon S, Gillis P, Horvitz E, Radinsky K. Gender-sensitive word embeddings for healthcare. J Am Med Inform Assoc. 2022;29:415-23.

[28] Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Commun Med (Lond). 2021;1:25.

[29] Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA. 2019;322:2377-8.

[30] Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Stat Med. 2012;31:2973-84.

[31] Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Spirit AI, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ. 2020;370:m3210.

[32] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26:1364-74.

[33] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368:m689.

[34] Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493-7.

[35] Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. 2020;3:147.

[36] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8:53.

[37] Liu H, Perl Y, Geller J. Transfer Learning from BERT to Support Insertion of New Concepts into SNOMED CT. AMIA - Annual Symposium proceedings. 2019:1129-38.

[38] Mulyar A, Uzuner O, McInnes B. MT-clinical BERT: scaling clinical information extraction with multitask learning. Am Med Inform Assoc. 2020;28:2108-15.

Appendix

The following Appendix includes additional information to improve the interpretability of the natural language processing articles.

Additional descriptive statistics for natural language processing datasets

Descriptive statistics for chapters that involved free-text analysis are included below. The chapters that included free-text analyses were Chapter 5, Chapter 6, Chapter 7, Chapter 8 and Chapter 9.

The included text in Chapter 5 had a median character length of 5129 and interquartile range 4232.75 to 6236.25. The mean character length was 5327.76 and standard deviation for character length 1490.68. The median word length was 730.5 and the interquartile range for word length was 607.25 to 897.75. The mean word length was 766.28 and the standard deviation of word length was 221.53. The total number of unique word stems in the dataset was 32335.

The included text in Chapter 6 had a median character length of 5472, and interquartile range 4595 to 6440.25. The mean character length was 5635.07 and standard deviation for character length 1458.55. The median word length was 768.5 and the interquartile range for word length was 627.5 to 914.5. The mean word length was 784.4 and the standard deviation of word length was 115.01. The total number of unique words in the dataset was 17481.

The included text in Chapter 7 had a median character length of 1711, and interquartile range 1284 to 2198. The mean character length was 1793.65 and standard deviation for character length 752.37. The median word length was 247 and the interquartile range for word length was 178 to 326. The mean word length was 269.42 and the standard deviation of word length was 171.91. The total number of unique words in the dataset was 21997.

The included text in Chapter 8 had a median character length of 1339, and interquartile range 926 to 1909. The mean character length was 1561.79 and standard deviation for character length 978.37. The median word length was 238 and the interquartile range for word length was 166 to 337. The mean word length was 272.94 and the standard deviation of word length was 157.99. The total number of unique words in the dataset was 161812.

The included text in the prospective validation dataset for Chapter 9 had a median character length of 1237, and interquartile range 880 to 1759. The mean character length was 1490.11 and standard deviation for character length 1013.34. The median word length was 167 and the interquartile range for word length was 115 to 244. The mean word length was 200.72 and the standard deviation of word length was 134.26. The total number of unique words in the dataset was 296406.

The included text in the external validation dataset for Chapter 9 had a median character length of 1292.5, and interquartile range 909 to 1824. The mean character length was 1454.28 and standard deviation for character length 777.26. The median word length was 177 and the interquartile range for word length was 120 to 255. The mean word length was 201.3 and the standard deviation of word length was 116.22. The total number of unique words in the dataset was 348657.

Examples of misclassifications in free-text analyses

The following examples are from Chapter 9. The first two examples are cases in which the model predicted discharge within the next 48 hours, but discharge did not occur in this timeframe. Examples 3 and 4 are cases in which the model predicted no discharge within the next 48 hours, but discharge did occur during this time period.

Example 1

General Medicine

S

Feels well today

No issues

Looks forward to leaving tomorrow

O

Hemodynamically stable + afebrile

Imp/ Medically stable

Plan

1. Respite tomorrow

2. Son would like to discuss lvl 4 package with SW - attempted 2x page

Example 2

GEN MED WR

# BACK PAIN

S/

Feeling well

Reports pain about 5-6 out of 10

Noticed given ibuprofen once during midnight

Nil use of PRN oxycodone in last 24hours

Nil further vomiting since yesterday

Concerns about mobility affected by back pain - worries that might not be able manage her pain at home

Agrees trial of regular analgesia

O/

RR 18 SpO2 97% on room air BP 154/71 HR 64 afebrile

BNO D4

Warm peripheries

Cap refill <2s

Pulse strong and regular

Moist mucous membrane

Chest clear anterolaterally

Bowel sounds present

Abdomen soft non tender

Calves soft non tender

Nil peripheral oedema

A/

Medically well and stable

Plan:

1. For regular ibuprofen

2. Await OT to organise 4WW

3. For regular aperients

- ensure bowel opens this PM

- microlax enema available PRN

4. For re-discuss with SW re: MAC referral

Example 3

Gen Med WR

# Multilobar CAP - on IV ceftriaxone, completed course of azithromycoin

# Asthma exacerbation secondary to CAP

# AF.

S/

Breathing feels unchanged today. TTE done today.

Explained likelihood of slow recovery given multilobar CAP.

Note event overnight - palpitations with chest pain that resolved with replacing O2

O/

RR 23, SpO2 96% on 4L

BP 100/65, HR 80

T 36

Warm well perfused.

JVP elevated

Chest - bronchial breath sounds in L upper and midzone. R scattered screps. Nil wheeze

Peripheral pitting oedema to mid shins bilaterally

A/

Tachypnoea and ongoing O2 requirement likely secondary to fluid overload and multilobar CAP (will be slow to improve)

P/

1. *Downgrade to PO amox/clav (7 days total Abx - last day Thu)*
2. *Commence furosemide 20mg daily*
3. *Chase TTE*
4. *Cease prednisolone after tomorrow*
5. *Encourage Mobility and sitting of bed as able.*
6. *Daily weight please.*
7. *Physio review for chest and mobility please.*
8. *May require period of respite on discharge to facilitate weaning O2*

Example 4

GEN MED WR

# Chronic back pain with degenerative changes

S/

reports eating and drinking well

Pain well controlled with Targin

Mobilising well

Reporting Bowels not open day 3

Happy to await Respite

O/E

Otherwise, patient is alert, appears well and sitting up on bed

Obs within normal limits

BNO D3

MMM

Chest clear

Abdo SNT

Calves SNT

Impression: Back pain improving with analgesia

P/

1. Continue analgesia

2. SW to continue discussion re: respite

3. Encourage oral intake and sitting out of bed /mobilise as tolerated

4. Regular lactulose and movicol charted to ensure bowel opening

5. Consider enema tomorrow if bowels not open by tommorow

Example of code

```
# code_example
# import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn
import scipy
import re
import nltk
import tensorflow as tf
import math
import xlrd
import xlwt
from sfunctions import get_preprocessed, get_model_lr, get_training, get_testing

# set params
param_proportion_test = 0.25
param_random_state = 10
param_inputs_selected = "note_section_content"
param_labels_selected = "2_day_stay"
param_filepath = 'filepath'
param_filename = 'filename'
param_ngram_range_lower = 1
param_ngram_range_upper = 3
```

```
param_max_features = 30000
param_k_folds = 5
param_stem = True
param_negatedetect = True
param_punctremov = True
param_removstopw = True
param_cutoff = 0.5

# load data
df = pd.read_csv(param_filepath + '\\' + param_filename)

# preprocess data
## encode outcome
df["2_day_stay"] = df["los_total"].dt.days
df["2_day_stay"] = df["2_day_stay"].mask(df["2_day_stay"] <= 1, 1)
df["2_day_stay"] = df["2_day_stay"].mask(df["2_day_stay"] > 1, 0)
df["7_day_stay"] = df["los_total"].dt.days
df["7_day_stay"] = df["7_day_stay"].mask(df["7_day_stay"] <= 6, 1)
df["7_day_stay"] = df["7_day_stay"].mask(df["7_day_stay"] > 6, 0)

## preprocess text
text_processed = get_preprocessed(df, param_inputs_selected, stem=param_stem,
                                  negatedetect=param_negatedetect,
                                  punctremov=param_punctremov,
                                  removstopw=param_removstopw)
```

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=param_max_features,
ngram_range=(param_ngram_range_lower,param_ngram_range_upper))
X = cv.fit_transform(text_processed).toarray()
Y = df[param_labels_selected]

## train_test_split
from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=param_proportion_test,
random_state=param_random_state)

# model
model = get_model_lr()

# training set
results_training = get_training(model, X_train, Y_train, kfolds=param_k_folds,
cutoff=param_cutoff)

# test set
results_test = get_testing(model, X_train, Y_train, X_test, Y_test, cutoff=param_cutoff)

# save results
dfresultstest = pd.DataFrame(results_test)
dfresultstest.to_excel('dfresutlstest.xlsx')
```

Examples of discharge summary templates

Discharge summary male

INTERIM PENDING CONSULTANT REVIEW

Mr #### stroke occurred on ####/####/####. He lives #### and works as ####. He is ####-hand dominant, mobilises #### and drives/does not drive. [Include other social history, ATSI status, residency status] His pre-stroke modified Rankin score was ####.

Key risk factors were ####. Other relevant past medical history includes ####.

Relevant pre-admission medications were: [e.g. antiplatelets, anticoagulants, incl. generic name, dose, frequency]

####

####

####

The patient was #### adherent.

His symptoms occurred while [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/####.

OR

He was last seen well [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/#### and his symptoms were noted at ####:#### on ####/####/####.

OR

He was well when he went to bed [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/#### and his symptoms were noted on awakening at ####:#### on

####/####/####.

His symptoms included #### [details of neurological deficit].

He presented to the RAH and a Code Stroke was #### activated. He was [alert/drowsy/obtunded/comatose]. His neurological deficits were #### and the NIHSS score was exactly/approximately ####. The patient could #### walk. He was in [cardiac rhythm] with a heart rate ####, blood pressure ####, temperature #### C and blood glucose level #### mmol/L. INR/other clotting assay was ####.
[Include any additional relevant information regarding physical examination/symptoms at presentation here]

A CT scan was performed at the RAH/FMC/LMH/other at ####:#### on ####/####/#### and demonstrated [insert key findings from CT scan; including any vessel occlusion sites; add information from CT perfusion if performed including DT3 lesion and CBF<30% core volume; include information regarding large vessel stenoses as relevant to stroke aetiology].

Intravenous thrombolysis was commenced at ####:#### on ####/####/#### with alteplase/tenecteplase [dose]. NIHSS immediately prior was #### or unchanged.
OR
Intravenous thrombolysis was not performed due to ####.

He proceeded to endovascular thrombectomy at ####:#### under conscious sedation/general anaesthesia with [retrieval device type] with groin/carotid/wrist puncture at ####:####. NIHSS immediately prior was #### [or unchanged]. A retrievable clot was identified in the

####/not identified. Reperfusion was #### achieved at ####:#### with resultant TICI #### flow. [add additional information regarding failed attempts, access site change (to wrist or carotid), multiple passes, stenting, antiplatelet or intra-arterial agent administration, anatomical issues, embolization into new territories or other intra-operative complication]

OR

He was deemed unsuitable for endovascular thrombectomy due to ####.

Reperfusion therapy was performed without complication.

OR

Reperfusion therapy was complicated by [e.g. sICH, other extracranial haemorrhage, angioedema; include time of onset, symptoms and subsequent management and outcome here.]

Follow-up MRI/CT was performed on (####/####/####) and revealed ####. 24-hour NIHSS score [if TPA/EVT] was ####.

[Insert any additional information relevant to admission and/or discharge and/or ongoing management]

Stroke aetiology work up:

- ECG (####/####/####): ####

- Holter monitor/Cardiac telemetry (####/####/#### to (####/####/####): ####

- Transthoracic echocardiogram (####/####/####): ####

- Transoesophageal echocardiogram (####/####/####): ####

- Carotid doppler ultrasound (####/####/####): ####

- Stroke related blood tests:

- HbA1c: ####%

- LDL ####, HDL ####, TGL ####

- Other [DSA, LP etc] (####/####/####): ####

Issues during his admission included:

# ####

[insert details DRPIMCO – diagnosis, risk factors, presenting symptoms, investigations, management, complications, outcome]

# ####

[insert details DRPIMCO – diagnosis, risk factors, presenting symptoms, investigations, management, complications, outcome]

[Audit collects data on DVT/PE/pressure ulcers/UTI/pneumonia/non-reperfusion therapy sICH (if sICH, date and time)]

His final diagnosis was a left/right/bilateral mild/moderate/severe ischaemic stroke/intracerebral haemorrhage [with or without intraventricular extension and hydrocephalus and estimated volume] affecting the #### cerebral artery territory/location due to ####.

He was referred back to FMC/LMH/other at ####:#### on ####/####/#### for #### care with an mRS of #### and a NIHSS of ####. Neurological examination at the time of transfer showed ####.

OR

He was discharged to home/[new] residential care/rehabilitation/other hospital/new residential care/relatives' home/residential TCP on ####/####/#### with an mRS of #### and a NIHSS of ####. Neurological examination at the time of discharge showed ####.

OR

He (was managed with palliative care from ####/####/#### and) died at ####:#### on ####/####/####.

Management plan:

1. #### [Medication changes and timing + reasons including addition of secondary stroke prevention measures]

2. #### [Follow up tests]

3. #### [Follow up appointments]

4. #### [Driving status/recommendations]

Thank you for your ongoing care.

____

Additional data for audit purposes:

Stroke onset:

- Location (#### postcode; home/RAH/other)

Transfer/transport

- Method of arrival: ambulance/walk-in/private car/hospital transfer/road retrieval/air

retrieval

- Hospital transfer: Y/N (hospital name and reason for transfer = possible tPA/ICU/endovascular Rx/Stroke Unit Care/neurosurgery/tertiary diagnostic tests; admission to primary hospital at ####:#### on ####/####/####; IV tPA prior to transfer = Y/N (type, #### on ####/####/####; telemedicine = Y/N)

- Ambulance arrival: SAAS (ID ####): time called ####:####, dispatched ####:####, arrived at patient ####:####, depart scene ####:####, arrive hospital ####:####; ambulance on bypass Y/N/unknown; ROSIER: Y/N (insert score) if >=1 ACT-FAST done? Y/N (positive/negative)

- Non-SAAS acute arrival : ROSIER: Y/N (insert score)

Pre-stroke risk factors and morbidity predictors: prior stroke (####/####/####, stroke subtype); prior TIA (####/####/####), obesity, hypertension, diabetes mellitus, dyslipidaemia, rheumatic heart disease, coronary heart disease, peripheral vascular disease, LVF/CHF, renal disease (eGFR ####), proteinuria, dialysis, Afib/flutter (prior to this stroke? Y/N; paroxysmal/permanent), active smoker, previous smoker, illicit drug use (details, incl. route), EtOH (#### standard drinks per day, binge pattern (>6 standard drinks), binge <24 hours prior), AIDS, active malignancy (details, metastases), dementia, liver disease (mild/mod/severe), peptic ulcer disease, DM microvascular complications, COPD, connective tissue disease, INR #### and date if on warfarin prior to presentation

Acute management data

- Admission to Stroke Unit: Y/N

- Antithrombotic commenced ####:#### on ####/####/####

- DVT prevention: LMWH/SCDs/therapeutic anticoagulation

Rehabilitation data

- Physio assessment ####:#### on ####/####/#### (contraindicated/patient refused/not documented),

- Treatment commenced by mobilisation (sitting/standing/walking) ####:#### on ####/####/#### (back to baseline/refused/comatose/palliated/no reason documented)

- A rehabilitation goal was #### documented

- Risk factor modification advice: Y/N (cognitive impairment/impaired communication/refused/palliative care/other/unknown);

- Carer assessment: Y/N (no carer/not discharged home/refused/complete recovery/not documented)

- Training Y/N (no carer/not discharged home/refused/complete recovery/not documented)

- Swallow screen: Y/N (####:#### on ####/####/####; #### passed); medication prior to swallow screen: Y/N

- Formal speech path assessment: Y/N (####:#### on ####/####/####)

- Stroke Care Plan: Y/N (####:#### on ####/####/####) (not discharged home/refused/not documented)

Discharge summary female

INTERIM PENDING CONSULTANT REVIEW

Ms #### stroke occurred on ####/####/####. She lives #### and works as ####. She is ####-hand dominant, mobilises #### and drives/does not drive. [Include other social history, ATSI status, residency status] Her pre-stroke modified Rankin score was ####.

Key risk factors were ####. Other relevant past medical history includes ####.

Relevant pre-admission medications were:

####

####

####

The patient was #### adherent.

Her symptoms occurred while [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/####.

OR

She was last seen well [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/#### and her symptoms were noted at ####:#### on ####/####/####.

OR

She was well when she went to bed [at home/in hospital/elsewhere] at exactly/approximately ####:#### on ####/####/#### and her symptoms were noted on awakening at ####:#### on ####/####/####.

Her symptoms included #### [details of neurological deficit].

She presented to the RAH and a Code Stroke was #### activated. She was [alert/drowsy/obtunded/comatose]. Her neurological deficits were #### and the NIHSS score was exactly/approximately ####. The patient could #### walk. She was in [cardiac rhythm] with a heart rate ####, blood pressure ####, temperature #### C and blood glucose level #### mmol/L. INR/other clotting assay was ####.
[Include any additional relevant information regarding physical examination/symptoms at presentation here]

A CT scan was performed at the RAH/FMC/LMH/other at ####:#### on ####/####/#### and demonstrated [insert key findings from CT scan; including any vessel occlusion sites; add information from CT perfusion if performed including DT3 lesion and CBF<30% core volume; include information regarding large vessel stenoses as relevant to stroke aetiology].

Intravenous thrombolysis was commenced at ####:#### on ####/####/#### with alteplase/tenecteplase [dose]. NIHSS immediately prior was #### or unchanged.
OR
Intravenous thrombolysis was not performed due to ####.

She proceeded to endovascular thrombectomy at ####:#### under conscious sedation/general anaesthesia with [retrieval device type] with groin/carotid/wrist puncture at ####:####. NIHSS immediately prior was #### [or unchanged]. A retrievable clot was identified in the ####/not identified. Reperfusion was #### achieved at ####:#### with resultant TICI #### flow. [add additional information regarding failed attempts, access site *change (to wrist or carotid), multiple passes, stenting, antiplatelet or intra-arterial agent administration, anatomical issues, embolization into new territories or other intra-operative complication]*

OR

She was deemed unsuitable for endovascular thrombectomy due to ####.

Reperfusion therapy was performed without complication.

OR

Reperfusion therapy was complicated by [e.g. sICH, other extracranial haemorrhage, angioedema; include time of onset, symptoms and subsequent management and outcome here.]

Follow-up MRI/CT was performed on (####/####/####) and revealed ####. 24-hour NIHSS score [if TPA/EVT] was ####.

[Insert any additional information relevant to admission and/or discharge and/or ongoing management]

Stroke aetiology work up:

- ECG (####/####/####): ####

- Holter monitor/Cardiac telemetry (####/####/#### to (####/####/####): ####

- Transthoracic echocardiogram (####/####/####): ####

- Transoesophageal echocardiogram (####/####/####): ####

- Carotid doppler ultrasound (####/####/####): ####

- Stroke related blood tests:

- HbA1c: ####%

- LDL ####, HDL ####, TGL ####

- Other [DSA, LP etc] (####/####/####): ####

Issues during his admission included:

# ####

[insert details DRPIMCO – diagnosis, risk factors, presenting symptoms, investigations, management, complications, outcome]

# ####

[insert details DRPIMCO – diagnosis, risk factors, presenting symptoms, investigations, management, complications, outcome]

[Audit collects data on DVT/PE/pressure ulcers/UTI/pneumonia/non-reperfusion therapy sICH (if sICH, date and time)]

Her final diagnosis was a left/right/bilateral mild/moderate/severe ischaemic stroke/intracerebral haemorrhage [with or without intraventricular extension and hydrocephalus and estimated volume] affecting the #### cerebral artery territory/location due to ####.

She was referred back to FMC/LMH/other at ####:#### on ####/####/#### for #### care with an mRS of #### and a NIHSS of ####. Neurological examination at the time of transfer showed ####.

OR

She was discharged to home/residential care/rehabilitation/other hospital/new residential care/relatives' home/residential TCP on ####/####/#### with an mRS of #### and a NIHSS of ####. Neurological examination at the time of discharge showed ####.

OR

She [was managed with palliative care from ####/####/#### and] died at ####:#### on ####/####/####.

Management plan:

1. #### [Medication changes and timing + reasons including addition of secondary stroke prevention measures]

2. #### [Follow up tests]

3. #### [Follow up appointments]

4. #### [Driving status/recommendations]

Thank you for your ongoing care.

____

Additional data for audit purposes:

Stroke onset:

- Location (#### postcode; home/RAH/other)

Transfer/transport

- Method of arrival: ambulance/walk-in/private car/hospital transfer/road retrieval/air retrieval

- Hospital transfer: Y/N (hospital name and reason for transfer = possible tPA/ICU/endovascular Rx/Stroke Unit Care/neurosurgery/tertiary diagnostic tests; admission to primary hospital at ####:#### on ####/####/####; IV tPA prior to transfer = Y/N (type, #### on ####/####/####; telemedicine = Y/N)

- Ambulance arrival: SAAS (ID ####): time called ####:####, dispatched ####:####, arrived at patient ####:####, depart scene ####:####, arrive hospital ####:####; ambulance on bypass Y/N/unknown; ROSIER: Y/N (insert score) if >=1 ACT-FAST done? Y/N (positive/negative)

- Non-SAAS acute arrival : ROSIER: Y/N (insert score)

Pre-stroke risk factors and morbidity predictors: prior stroke (####/####/####, stroke subtype); prior TIA (####/####/####), obesity, hypertension, diabetes mellitus, dyslipidaemia, rheumatic heart disease, coronary heart disease, peripheral vascular disease, LVF/CHF, renal disease (eGFR ####), proteinuria, dialysis, Afib/flutter (prior to this stroke? Y/N; paroxysmal/permanent), active smoker, previous smoker, illicit drug use (details, incl. route), EtOH (#### standard drinks per day, binge pattern (>6 standard drinks), binge <24 hours prior), AIDS, active malignancy (details, metastases), dementia, liver disease (mild/mod/severe), peptic ulcer disease, DM microvascular complications, COPD, connective tissue disease, INR #### and date if on warfarin prior to presentation

Acute management data

- Admission to Stroke Unit: Y/N

- Antithrombotic commenced ####:#### on ####/####/####

- DVT prevention: LMWH/SCDs/therapeutic anticoagulation

Rehabilitation data

- Physio assessment ####:#### on ####/####/#### (contraindicated/patient refused/not documented),

- Treatment commenced by mobilisation (sitting/standing/walking) ####:#### on ####/####/#### (back to baseline/refused/comatose/palliated/no reason documented)

- A rehabilitation goal was #### documented

- Risk factor modification advice: Y/N (cognitive impairment/impaired communication/refused/palliative care/other/unknown);

- Carer assessment: Y/N (no carer/not discharged home/refused/complete recovery/not documented)

- Training Y/N (no carer/not discharged home/refused/complete recovery/not documented)

- Swallow screen: Y/N (####:#### on ####/####/####; #### passed); medication prior to swallow screen: Y/N

- Formal speech path assessment: Y/N (####:#### on ####/####/####)

- Stroke Care Plan: Y/N (####:#### on ####/####/####) (not discharged home/refused/not documented)