Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing

Anthony Shek, Zhilin Jiang, James Teo, Joshua Au Yeung, Ajay Bhalla, Mark P. Richardson, Yee Mah

Abstract

Background and purpose: With the increasing adoption of electronic records in the health system, machine learning-enabled techniques offer the opportunity for greater computer-assisted curation of these data for audit and research purposes. In this project, we evaluate the consistency of traditional curation methods used in routine clinical practice against a new machine learning-enabled tool, MedCAT, for the extraction of the stroke comorbidities recorded within the UK’s Sentinel Stroke National Audit Programme (SSNAP) initiative.

Methods: A total of 2327 stroke admission episodes from three different National Health Service (NHS) hospitals, between January 2019 and April 2020, were included in this evaluation. In addition, current clinical curation methods (SSNAP) and the machine learning-enabled method (MedCAT) were compared against a subsample of 200 admission episodes manually reviewed by our study team. Performance metrics of sensitivity, specificity, precision, negative predictive value, and F1 scores are reported.

Results: The reporting of stroke comorbidities with current clinical curation methods is good for atrial fibrillation, hypertension, and diabetes mellitus, but poor for congestive cardiac failure. The machine learning-enabled method, MedCAT, achieved better performances across all four assessed comorbidities compared with current clinical methods, predominantly driven by higher sensitivity and F1 scores.

Conclusions: We have shown machine learning-enabled data collection can support existing clinical and service initiatives, with the potential to improve the quality and speed of data extraction from existing clinical repositories. The scalability and flexibility of these new machine-learning tools, therefore, present an opportunity to revolutionize audit and research methods.

Keywords: clinical coding, natural language processing, programme evaluation, stroke
INTRODUCTION

Medical records are a rich source of information, continuously accessed by health care professionals to help care for their patients and community. The benefits of trawling through swathes of medical notes are clear, including understanding the individual in the acute setting; audit and service evaluation [1–3]; and identifying patterns embedded in a disease population for research [4–6]. With the increasing adoption of electronic records in the health system [7–10], using computers to analyse all these data has been a common objective [11–13]. However, accurate extraction of medical concepts from unstructured data, like free text, requires an understanding of the language used, something that is relatively simple for a human but extremely challenging for a computer.

Over the past decade advancements in a branch of machine learning, known as natural language processing (NLP), have enabled the translation of free text into a standardized, structured set of medical terms that can be subsequently analysed by a computer [14]. These tools have the potential to automate and support data collection; however, evaluation with real-world clinical data across medical specialties, such as stroke, has been limited [15]. CogStack is an open-source software ecosystem that incorporates both the structured and unstructured components of the electronic health record (EHR). The MedCAT Toolkit [16] component supports the development of NLP algorithms through the ability to disambiguate and capture synonyms, and acronyms for medical Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) concepts, as well as surrounding linguistic context such as negation, subject, and basic grammatical tense, using deep learning and long short-term memory networks. Further supervised training can improve the detection of annotations and meta-annotations using the MedCAT Trainer platform. The entire CogStack ecosystem is open source and available on GitHub (https://github.com/CogStack).

Conventional registry and national audits use standardized case report forms to provide periodic standardized submissions into centralized databases. The Sentinel Stroke National Audit Programme (SSNAP) [17] is a health care quality improvement programme collecting stroke patient data that represent >90% of all cases in England, Wales, and Northern Ireland. With 100,000 stroke cases per annum [17], this is a time-pressured, labour-intensive exercise conducted manually by a team of clinical coders and/or clinicians. Although manual curation is the current gold standard, these pressures increase the risk of errors [18–20] and limit the timeliness of the data to some months after the event, negatively impacting on the utility of the collected data.

In this project, we evaluated the manually inputted SSNAP data from three different National Health Service (NHS) hospitals against a manually reviewed sample for the four stroke comorbidities routinely collected as part of the SSNAP initiative. We also trained MedCAT on a set of manually annotated stroke documents, to identify the same four comorbidities, and then applied the model to the inpatient stroke records of three different NHS hospitals, comparing the MedCAT performances against the corresponding manually inputted SSNAP data and the manually reviewed subsample.

METHODS

Datasets

All admission episodes contained within SSNAP were identified for the following three NHS hospitals in the time period 1 January 2019 to 1 April 2020: King’s College Hospital NHS Foundation Trust (KCH); Princess Royal University Hospital NHS Trust (PRUH); and Guy’s and St Thomas’ NHS Foundation Trust (GSTT). KCH and PRUH share a common EHR system, whereas GSTT uses a clinically separate electronic documentation system. A total of 2327 admission episodes were included in this study. Table 1 shows the breakdown of episodes. Patients without any available electronic notes were excluded from the study.

This project was conducted under audit and data processing for service evaluation. Research ethics review was not required. For

### TABLE 1 Table showing the distribution of patient episodes and comorbidities for the different datasets

|                | KCH-PRUH | KCH subsample | p | GSTT | GSTT subsample | p |
|----------------|----------|---------------|---|------|----------------|---|
| Number of episodes | 2136     | 100           |   | 191  | 100            |   |
| Excluded episodes  | 124      | –             |   | 18   | –              |   |
| Female            | 1066 (49.9%) | 51           |   | 96 (50.3%) | 47            |   |
| Mean age, years   | 71.5 (SD = 15.1) | 66.9 (SD = 16.0) | 0.58 | 70.9 (SD = 14.4) | 71.3 (SD = 12.5) | 0.94 |
| AIS               | 1864 (87.3%) | 88 (88%)      | 0.52 | 166 (86.9%) | 89 (89%)      | 0.08 |
| AF                | 446 (20.9%) | 21 (21%)      | 0.88 | 35 (18.3%) | 15 (15%)      | 0.79 |
| Hypertension      | 1235 (57.6%) | 60 (60%)      | 0.97 | 109 (57.1%) | 62 (62%)      | 0.91 |
| CCF               | 90 (4.2%)  | 8 (8%)        | 0.08 | 12 (6.3%) | 7 (7%)        | 0.40 |
| Diabetes          | 520 (24.3%) | 32 (32%)      | 0.63 | 59 (30.1%) | 32 (32%)      | 0.44 |

Note: Patient episodes without digital documents were excluded. A Wilcoxon rank-sum test was performed to assess for differences between the subsample and parent sample for KCH-PRUH and GSTT.

Abbreviations: AF, atrial fibrillation; AIS, acute ischaemic stroke; CCF, congestive cardiac failure; GSTT, Guy’s and St Thomas’ Hospital; KCH, King’s College Hospital; PRUH, Princess Royal University Hospital.
MedCAT algorithm training and data

The base MedCAT algorithm was trained in an unsupervised manner on the entire KCH EHR data consisting of more than 18 million documents [16], and received additional training from 301 and 373 annotated documents in endocrinology and cardiology, respectively. For our study, further training on stroke-specific comorbidities was provided through 500 KCH-PRUH annotated documents obtained from 2015 to October 2020, stratified by patient, age, and gender, using the method described in Kraljevic et al. [16]. This only included free-text information documented by clinical staff, and excluded information from other systems like blood results, investigation reports, outpatient letters, and vital observations. MedCAT counted the number of instances a concept was mentioned (e.g., atrial fibrillation [AF]) and generated a total count for each patient episode. This only included references for the presence of the concept relating to a patient. Phrases such as “this patient does not have AF” or “a family history of AF” would not increase the count. Because MedCAT is mapped onto the SNOMED-CT library, counts for child concepts defined by the inbuilt “IS A” hierarchical relationship were merged to reflect the SNOMED-CT concepts used to emulate the SSNAP comorbidities.

The MedCAT concept count was converted to a binary state by applying a threshold, above which a patient would be diagnosed with the comorbidity for the specific admission episode. Two different document periods were examined, based on the recorded admission and discharge timestamps: (i) 12 h prior to admission to 12 h after discharge (admission period) and (ii) January 2015 to 12 h after discharge (2015-to-discharge).

SSNAP data

The SSNAP governing body has released protocols and guidelines for data curation, with each participating site responsible for its own curation of data [17]. Although the data collected for SSNAP have evolved with the changing face of stroke, the data definition for SSNAP remained constant during the period assessed in our study.

Using the local SSNAP data from each hospital, the comorbidities AF, hypertension, congestive cardiac failure (CCF), and diabetes were extracted. SSNAP collects both “atrial fibrillation” and “new atrial fibrillation,” where the patient cannot be positive for both labels. For this project, we combined these two groups to represent whether the comorbidity of AF was present for this admission episode and used this to perform the subsequent analyses. Diabetes included both Type 1 and Type 2 diabetes mellitus. In addition, the stroke type for the admission episode—acute ischaemic stroke (AIS) or primary intracerebral haemorrhage (PIH)—was recorded.

Subsample reread (KCH and GSTT): Ground truth

To evaluate the performance of the two auditing methods, a mutual reference dataset was curated to represent the ground truth. Two subsamples of 100 patient episodes each were randomly selected from the KCH and GSTT datasets. A Wilcoxon rank-sum test was used to assess whether there were significant differences between the subsample reread and its parent dataset.

To curate the subsample reread, the range of documents and level of access were identical to what would have been available to the SSNAP operators. In contrast to the SSNAP curation method, the subsample reread was collected solely by trained clinicians (Y.M., J.T., A.B., Z.J., J.A.Y.) who were not under time constraints to extract the comorbidities. The medical notes for each admission episode were reviewed, excluding any documents that were created after the episode discharge date. The presence or absence of AF, hypertension, CCF, and diabetes was noted. They also recorded whether the acute admission was either an AIS or a PIH. For each site, the level of interrater agreement for the four comorbidities was assessed using Cohen kappa on 20% of the sample. In cases of disagreement between operators, the finding from the more senior clinician was used.

Evaluation

For purposes of brevity, SSNAP is used to indicate data held in SSNAP that are obtained through manual data curation. MedCAT is used to indicate data generated by NLP from hospital source documents. A series of comparisons were then performed to evaluate the two auditing methods using the three different datasets: MedCAT, SSNAP, and subsample reread (Figure 1).

Two sets of comparisons were conducted: first, MedCAT was assessed against SSNAP data for the two different EHR systems (Figure 1, comparison A); second, both MedCAT and SSNAP were assessed against the ground truth, subsample reread of KCH (n = 100) and GSTT (n = 100) datasets (Figure 1, comparisons B and C). Metrics of sensitivity, specificity, precision, negative predictive value, and F1 score were calculated. The level of agreement between the subsample reread and the auditing methods was measured using Cohen kappa, whereas the McNemar test was used to assess the difference in performance between the auditing methods. All statistical analyses were performed using MATLAB 2020b [22].
The mean ages for the KCH-PRUH and GSTT datasets were 71.5 and 70.9 years, respectively. There was no significant difference between the subsample reread and the respective parent dataset for age, proportion of females, AIS, and comorbidities. The absolute values for each dataset are displayed in Table 1. There was a significantly higher proportion of females at the GSTT site compared with KCH-PRUH ($p = 0.045$). The prevalence of comorbidities between the KCH-PRUH and GSTT sites was not significantly different except for diabetes mellitus ($p = 0.045$).

The F1 score is the harmonic mean between the sensitivity and precision (positive predictive value). For the KCH-PRUH and GSTT datasets, compared against SSNAP, MedCAT was able to determine the type of stroke using documents from 2015-to-discharge with peak F1 scores of 0.92 and 0.95, respectively (Table S2).

**Comparison A: MedCAT compared against SSNAP**

Comparing MedCAT against SSNAP, the performance of MedCAT is similar between the two document inclusion periods, with the peak F1 scores obtained within a threshold value between two and eight counts (Table S1). MedCAT’s performance as a function of MedCAT count threshold for the type of stroke, and the four comorbidities are provided in Figures S1 and S2. The corresponding area under the receiver operating characteristic curve plots for the four comorbidities are displayed in Figure S3.

The peak F1 scores for AF and diabetes were obtained using documents from the admission period only, whereas CCF was from all documents (2015-to-discharge), and there was no difference with hypertension (Table S1). The deterioration in F1 score for AF in the GSTT data compared with the KCH-PRUH data is primarily driven by a low level of precision (i.e., false positives) likely related to the number of acronyms for atrial fibrillation (e.g., "AF," "PAF," "AFib," "atrial fibr") not encountered in the training sample.

**Comparison B and C: MedCAT/SSNAP compared against the subsample reread (ground truth)**

Interrater agreement for the two sites were high, with a Cohen kappa of 0.89 for diabetes, and 1.0 for the other three comorbidities at the KCH site, and 0.83 for hypertension and CCF, and 1.0 for AF and diabetes at the GSTT site.

To facilitate the comparison, a threshold heuristic of five counts was selected for the MedCAT models used at each site and applied to document inclusion periods of admission only and
2015-to-discharge. In the KCH dataset, SSNAP obtained F1 scores of 0.84, 0.91, and 0.91 for AF, hypertension, and diabetes, with CCF performing worse, with a score of 0.54. In comparison, MedCAT achieved F1 scores greater than or equal to 0.91 in all but CCF, with a score of 0.71 (Table 2). The F1 scores for MedCAT were higher when using all available documents until discharge (2015-to-discharge; Table S2). Similarly, in the GSTT dataset, SSNAP obtained F1 scores of 0.67, 0.81, and 0.85 for AF, hypertension, and diabetes, with a lower score of 0.40 for CCF (Table 2). MedCAT obtained F1 scores greater than 0.8 for all comorbidities apart from CCF, which was 0.64. Unlike the KCH dataset, the F1 scores were not consistently higher when using all available documents, compared with only those from the admission period, with the reverse found for AF and diabetes (Table S2).

The MedCAT audit method achieved substantial levels of agreement with the subsample reread at both sites (Table 3), with near-perfect agreement achieved with AF at KCH, and diabetes at both sites. In contrast, SSNAP achieved a lower level of agreement for all comorbidities at both sites, except for hypertension at KCH. The MedCAT audit method was significantly superior to SSNAP for hypertension in the GSTT dataset, and almost reached significance for CCF in both the KCH and GSTT datasets, with a p-value of 0.051 and 0.079, respectively, based on the McNemar test. The percentage of agreement between the MedCAT and SSNAP auditing methods when evaluated against the subsample reread is presented in Table S4.

**DISCUSSION**

The SSNAP is a national health care quality improvement programme supporting the delivery of evidence-based care for stroke. It has helped shape stroke services in the UK by measuring process of care, with data collected under time pressure in a continuous and contemporaneous manner. A hub-and-spoke system exists in the UK, where a hyperacute stroke unit (HASU) provides hyperacute intervention and care to a large area containing multiple smaller stroke units (SUs) that manage longer term rehabilitation needs. With a high annual incidence of stroke, the task of data collection is shared, with each hospital required to submit data for every patient that passes through its unit. In this project, we have examined the four comorbidities recorded by SSNAP over a 15-month period and evaluated the consistency of current audit practices, as well as a new machine learning-enabled method, MedCAT, against a manually reviewed set of patient episodes.

To evaluate an auditing method, a ground truth needs to be specified, and will inevitably require human operator involvement. Here, we have referred to our "subsamp[le reread]" as the ground truth and compared both the MedCAT and SSNAP methods against this dataset. Although this subsample reread is potentially vulnerable to the same risks of human error as in the SSNAP method, it was importantly not encumbered by the same time pressures that afflict SSNAP, focussed on comorbidities alone, and was performed completely by clinicians who would be better able to interpret the medical vernacular and extract the appropriate concepts. Moreover, the interrater consistency for both sites were strong, with a Cohen kappa greater than 0.80 for all comorbidities.

The KCH and PRUH sites operate both an HASU and an SU, with more than 1800 patient episodes annually between them. There was good consistency between SSNAP and our subsample reread for AF, hypertension, and diabetes. The F1 score for CCF was poor, primarily driven by low sensitivity. This is partly explained by the low prevalence of CCF within the subsample reread, with small numbers of detection errors incurring a large deterioration in performance. Importantly, the prevalence of heart failure in a nonenriched population aged more than 55 years is 7% [23], whereas the proportion of patients with CCF in the entire KCH-PRUH SSNAP data was 4.2%, therefore suggesting underrepresentation. As a clinical syndrome, diagnosis relies on the recognition of symptoms and signs, in combination with the interpretation of cardiac investigations to ensure accurate heart failure classification [24,25]. Consequently, diagnosing CCF from the clinical notes may be more challenging compared with the other comorbidities, with more junior clinical staff less likely to document the diagnosis explicitly or using poecilonyms like "heart failure" or "pulmonary oedema from heart failure." Although the diagnosis may be recorded in other documents, such as clinic letters, these may not be available to the acute team, especially if these documents originate from a different trust due to the organization of stroke services. In this scenario, data availability will rely on the recall of the patient and relative. A similar pattern is obtained from the data from GSTT, which only operates an SU. This operational

---

**Table 3** Table showing the performance metrics of SSNAP and MedCAT against the manually reviewed (ground truth) subset

|        | SSNAP, Cohen kappa (95% CI) | MedCAT, Cohen kappa (95% CI) | McNemar p-value |
|--------|----------------------------|------------------------------|-----------------|
| KCH    |                            |                              |                 |
| AF     | 0.79 (0.65–0.93)           | 0.86 (0.75–0.97)             | 0.296           |
| HTN    | 0.74 (0.60–0.88)           | 0.69 (0.53–0.84)             | 0.583           |
| CCF    | 0.48 (0.21–0.76)           | 0.67 (0.45–0.89)             | 0.051           |
| DM     | 0.87 (0.77–0.97)           | 0.90 (0.81–0.98)             | 0.369           |
| GSTT   |                            |                              |                 |
| AF     | 0.60 (0.41–0.80)           | 0.64 (0.48–0.80)             | 0.295           |
| HTN    | 0.51 (0.33–0.70)           | 0.75 (0.60–0.91)             | <0.001          |
| CCF    | 0.51 (0.19–0.84)           | 0.64 (0.35–0.92)             | 0.079           |
| DM     | 0.76 (0.63–0.89)           | 0.86 (0.75–0.96)             | 0.186           |

Note: The MedCAT model used a threshold heuristic of 5 and was applied to all available documents until discharge (2015-to-discharge). The level of agreement between the ground truth subsample reread and the auditing method was calculated using Cohen kappa, whereas McNemar test was applied to assess whether MedCAT classified patients more accurately than SSNAP (p-values are provided).

Abbreviations: AF, atrial fibrillation; CCF, congestive cardiac failure; DM, diabetes mellitus; GSTT, Guy’s and St Thomas’ Hospital; HTN, hypertension; KCH, King’s College Hospital; SSNAP, Sentinel Stroke National Audit Programme.
difference may explain the apparent deterioration in F1 score for AF, as the inpatient documentation is likely to have a greater emphasis on rehabilitation requirements rather than the aetiology of the stroke.

The low sensitivity of the SSNAP data compared with the subsample reread at KCH and GSTT would indicate sufficient information is present within the documents available at each site. The absence of time pressure, and requirement to extract only the comorbidities rather than all the SSNAP concepts, are likely to have contributed to greater sensitivity in the subsample reread, especially when a more comprehensive review of the records is required.

MedCAT uses NLP to extract concepts from the free text and maps them onto a standardized clinical vocabulary, SNOMED-CT. Intuitively, a more accurate picture of patients’ stroke risk profiles will be obtained from a review of their entire medical histories. Clinical teams will review a patient’s extensive history to identify potentially relevant stroke risk factors using sources from within and external to the hospital. This phenomenon is demonstrated in the performance of MedCAT, with higher F1 scores obtained when using all available documents (Table S2), although the effect appears less pronounced with the GSTT dataset. The machine learning-driven process of MedCAT achieved a higher F1 score for all comorbidities than the manually curated approach adopted by SSNAP. This increase in consistency is secondary to an improvement in sensitivity, with the superior performance of the machine learning-enabled approach reaching significance for CCF in the GSTT subsample reread. This therefore explains the lower F1 scores achieved by MedCAT when comparing against the SSNAP dataset, as it is simply reflecting the inherent errors of SSNAP.

There are several limitations to this audit. First, MedCAT was restricted to documents dated after January 2015, and did not include clinic letters or investigation reports, thereby reducing the sensitivity of the method. Modifying the documents available to MedCAT is possible and may improve sensitivity, particularly for “congestive cardiac failure” detection, especially if there are several closely associated but nonsynonymous concepts like “pulmonary oedema” and “heart failure.” Second, the MedCAT model used to extract the comorbidities was initially trained on 500 manually annotated documents from KCH and PRUH electronic notes, thereby exposing the model to some of the data prior to this audit. The effect of this is likely to be small, as it represents 0.08% of the total number of documents. Importantly, there was no training performed on the GSTT data, with a similar level of performance achieved, highlighting the feasibility of using a validated general model across different sites. Further training of the model on local hospital documents could then be subsequently performed to optimize MedCAT to the local environment. Third, MedCAT requires the electronic medical notes to be centralized into a searchable state prior to execution. Although MedCAT and CogStack are both open source, setup of these systems can require significant upfront investment, depending on the size of the project, and may be considered a limiting factor to wider adoption. However, this needs to be considered against the running costs of a physician’s time, and of a potential dedicated full-time curator. Fourth, despite the enriched population, the prevalence of the comorbidities is low, except for hypertension. Consequently, the training of the MedCAT model to the stroke-specific concepts was biased toward those concepts with greater representation within the population. It is unsurprising to see both AF and CCF perform worse than hypertension and diabetes. This is highlighted with errors where MedCAT misattributes the abbreviation “AF” to “atrial fibrillation” rather than “artificial feed” despite the nutritional context of the entry. Nevertheless, this issue can be addressed with further focussed training or more sophisticated NLP models (e.g., transformer-based models).

The UK SSNAP initiative requests participating centres to submit data within set deadlines to facilitate analysis and reporting. Each centre will utilize the existing clinical team, and may, as is the case at our three NHS sites, employ a dedicated trained operator to manually review the medical notes and extract the necessary concepts. This is representative of the arrangements across London, and likely throughout the rest of the United Kingdom. Our audit of stroke comorbidities collected by SSNAP shows reasonable performance of current methods, except for CCF, when compared with our subsample reread. The inconsistencies between the two datasets comprise a combination of factors, from limited physician involvement to time constraints in its collection. With the advancements in NLP, we have shown that machine learning-enabled data collection, in the form of MedCAT, has at least comparable concept extraction to the traditional manual processes used for SSNAP, and may potentially improve the quality of data when compared against our manual subsample.

MedCAT has several advantages. First, the speed of data collection is many orders of magnitude faster than manual review. With the CogStack ecosystem and MedCAT established within the hospital, patients’ notes can be continuously monitored, providing near real-time data extraction if desired. Second, MedCAT is capable of interrogating huge volumes of data, applying the same level of scrutiny throughout the entire medical records, whether the patient has a short or extensive medical history. Third, deployment of different models is relatively simple. We have shown that a MedCAT model naive to the local documents is noninferior to traditional methods employed by SSNAP. Additionally, it possesses the flexibility to adapt existing models through further training to address other medical and service questions. Moreover, deployment is primarily digitally delivered, potentially providing a more responsive and scalable system.

In conclusion, we have shown that machine learning-enabled data collection can support existing clinical and service initiatives, potentially improving the quality and speed of data extraction from existing clinical repositories. Clearly, extracting information located within densely packed resources is challenging for humans, especially medical notes, which are not structured, but are evolving documents, consisting of thoughts from multiple users over time. The scalability and flexibility of MedCAT therefore present an opportunity to revolutionize audit and research methods.
ACKNOWLEDGMENTS

The supporting infrastructure and code base received funding from NIHR Maudsley BRC, Health Data Research UK, UK Research and Innovation, London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare, Innovate UK, the NIHR Applied Research Collaboration South London, Office of Life Sciences (UK), and NHSX.

CONFLICT OF INTEREST

J.T. has received research grant funding support from Innovate UK, NHSX, Office of Life Sciences, Bristol-Meyers Squibb, and Pfizer; speaker honoraria from Bayer, Bristol-Meyers Squibb, Pfizer, and Goldman Sachs; hospitality from iRhythm Technologies; and copyright fees from Wiley-Blackwell; and owns public shares in Nvidia, Amazon, and Alphabet.

AUTHOR CONTRIBUTIONS

Anthony Shek: Data curation (equal), formal analysis (equal), writing-review & editing (equal). Zhilin Jiang: Data curation (equal), writing-review & editing (equal). James Teo: Data curation (equal), formal analysis (equal), supervision (equal), writing-review & editing (equal). Joshua Au Yeung: Data curation (equal), writing-review & editing (equal). Ajay Bhalla: Data curation (equal), writing-review & editing (equal). Mark P. Richardson: Supervision (equal), writing-review & editing (equal). Yee Mah: Conceptualization (equal), data curation (equal), formal analysis (equal), supervision (equal), writing-review & editing (equal).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

James Teo https://orcid.org/0000-0002-6899-8319
Yee Mah https://orcid.org/0000-0002-6652-9922

REFERENCES

1. Dutta D, Thornton D, Bowen E. Using population-based routinely collected data from the Sentinel Stroke National Audit Programme to investigate factors associated with discharge to care home after rehabilitation. Clin Rehabil. 2018;32(8):1108-1118.
2. Xu X-M, Vestesson E, Paley L, et al. The economic burden of stroke care in England, Wales and Northern Ireland: Using a national stroke register to estimate and report patient-level health economic outcomes in stroke. Eur Stroke J. 2018;3(1):82-91.
3. Great Britain, National Audit Office, Morse A. Department of Health: progress in improving stroke care. London: The Stationery Office; 2010.
4. Melamed RD, Khbianian H, Rabadan R. Data-driven discovery of seasonally linked diseases from an Electronic Health Records system. BMC Bioinformatics. 2014;15(Suppl 6):S3.
5. Bean DM, Kraljevic Z, Searle T, et al. Angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers are not associated with severe COVID-19 infection in a multi-site UK acute hospital trust. Eur J Heart Fail. 2020;22(6):967-974.
6. Mah Y-H, Nachev P, MacKinnon AD. Quantifying the impact of chronic ischemic injury on clinical outcomes in acute stroke with machine learning. Front Neurol. 2020;11:15. [cited 2020 May 9]. 10.3389/fneur.2020.00151/full
7. NHS England, Local Digital Roadmaps, Personalised Health and Care 2020. Using Data and Technology to Transform Outcomes for Patients and Citizens: A framework for Action [Internet]. Department of Health and Social Care; 2014. Available from: https://www.gov.uk/government/publications/personalised-health-and-care-2020-using-data-and-technology-to-transform-outcomes-for-patients-and-citizens
8. Electronic Health Records. Parliamentary Office of Science and Technology; 2016 Feb. Report No.: 519.
9. Office-based Physician Electronic Health Record Adoption [Internet]. Office of the National Coordinator for Health Information Technology - A division of the U.S. Department of Health and Human Services; 2019. Available from: dashboard.healthit.gov/quickstats/pages/physician-ehr-adoptions-trends.php
10. WHO Global Observatory for eHealth, ProQuest (Firm). Global diffusion of eHealth: making universal health coverage achievable: report of the third global survey on eHealth [Internet]; 2016 [cited 2021 May 23]. Available from: https://ebookcentral.proquest.com/lib/quot/detail.action?docID=5910090
11. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc. 2008;15(1):36-39.
12. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical. in: proceedings of the 2nd clinical natural language processing workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019 [cited 2021 May 23]. p. 72-78. Available from: http://aclweb.org/anthology/W19-1909
13. SI Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc. 2019;26(11):1297-1304.
14. Fu S, Chen D, He H, et al. Clinical concept extraction: a methodology review. J Biomed Inform. 2020;109:103526.
15. Wang Y, Wang L, Rastegar-Mojarrad M, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34-49.
16. Kraljevic Z, Searle T, Shek A, et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. Artif Intell Med. 2021;117:102083.
17. Royal College of Physicians, Clinical Effectiveness and Evaluation Unit on behalf of the Intercollegiate Stroke Working Part. Royal College of Physicians Sentinel Stroke National Audit Programme (SSNAP). National clinical audit annual results portfolio. March 2016-April 2017 [Internet]. Royal College of Physicians; 2017. Available from: http://bit.ly/1NHYlqH Last accessed October 24, 201
18. The quality of clinical coding in the NHS [Internet]. Capita; 2014. Available from: https://assets.publishing.service.gov.uk/govermment/uploads/system/uploads/attachment_data/file/364476/The_quality_of_clinical_coding_in_the_NHS.pdf
19. Santos S, Murphy G, Baxter K, Robinson KM. Organisational factors affecting the quality of hospital clinical coding. Health Inf Manag J. 2008;37(1):25-37.
20. Alonso V, Santos JV, Pinto M, et al. Problems and barriers during the process of clinical coding: a focus group study of coders’ perceptions. J Med Syst. 2020;44(3). [cited 2021 May 23] 10.1007/s10916-020-1532-x
21. National Data Guardian. Guidance: The Caldicott Principles [Internet]; 2020. Available from: https://www.gov.uk/government/publications/the-caldicott-principles
22. MATLAB Version 2020b. Matick, MA: The MathsWorks Inc; 2020. Computer Software.
23. Bleumink GS, Knetsch AM, Sturkenboom MCJM, et al. Quantifying the heart failure epidemic: prevalence, incidence rate, lifetime risk and prognosis of heart failure The Rotterdam Study. *Eur Heart J.* 2004;25(18):1614-1619.

24. Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur Heart J.* 2016;37(27):2129-2200.

25. Yancy CW, Jessup M, Bozkurt B, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol.* 2013;62(16):e147-239.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

Figure S1
Figure S2

How to cite this article: Shek A, Jiang Z, Teo J, et al. Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. *Eur J Neurol.* 2021;28:4090-4097. [https://doi.org/10.1111/ene.15071](https://doi.org/10.1111/ene.15071)