Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings

Anthony N. Nguyen, PhD[1], Donna Truran[1], Madonna Kemp[1], Bevan Koopman, PhD[1], David Conlan[1], John O'Dwyer[1], Ming Zhang[1], Sarvnaz Karimi, PhD[2], Hamed Hassanzadeh, PhD[1], Michael J. Lawley, PhD[1], Damian Green[3]
[1]The Australian e-Health Research Centre, CSIRO, Brisbane/Sydney/Perth, Australia; [2]Data61, CSIRO, Sydney, Australia; [3]Gold Coast Hospital and Health Service, Department of Health, Queensland Government, Gold Coast, Australia

Abstract

Computer-assisted (diagnostic) coding (CAC) aims to improve the productivity and accuracy of clinical coders. The level of accuracy, especially for a wide range of complex and less prevalent clinical cases, remains an open research problem. This study investigates this problem on a broad spectrum of diagnostic codes and, in particular, investigates the effectiveness of utilising SNOMED CT for ICD-10 diagnosis coding. Hospital progress notes were used to provide the narrative rich electronic patient records for the investigation. A natural language processing (NLP) approach using mappings between SNOMED CT and ICD-10-AM (Australian Modification) was used to guide the coding. The proposed approach achieved 54.1% sensitivity and 70.2% positive predictive value. Given the complexity of the task, this was encouraging given the simplicity of the approach and what was projected as possible from a manual diagnosis code validation study (76.3% sensitivity). The results show the potential for advanced NLP-based approaches that leverage SNOMED CT to ICD-10 mapping for hospital in-patient coding.

Introduction

Australian hospitals invest in clinical coders to abstract relevant information from patients' medical records and decide which diagnoses and procedures meet the criteria for coding as per Australian Coding Standards[1]. The assigned codes and other patient data are used to determine a Diagnosis Related Group (DRG) for the episode of care, which is used for funding and reimbursement. Other common uses of the coded data include clinical research and audits, health services planning and resource allocation, epidemiological studies, benchmarking and education.

Clinical coders translate information from patient medical records into alphanumeric codes. The International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification (ICD-10-AM)[*], is applied in all Australian acute health facilities. For example, acute appendicitis is represented by the ICD-10-AM code 'K35.8'.

Clinical coding is a specialized skill requiring excellent knowledge of medical terminology, disease processes, and coding rules, as well as attention to detail and analytical skills[1]. The process mainly relies on manual inspections and experience-based judgments from clinical coders, and the effort required for information abstraction is extremely labor and time intensive and prone to human errors. Inaccuracies in coding can result in significant missed revenue[2]. As a result, there is an increasing demand for experienced and accurate clinical coders but at the same time an increasing lack of supply[3]. Strategies and opportunities for improving clinical coding productivity are thus required.

Advances in computer-assisted coding (CAC) may reduce demand for clinical coders. Automatic generation of codes from a CAC system for coder review and validation can improve the productivity of clinical coders[4]. Despite reports of streamlining processes, improving coding efficiencies and coding accuracy, current solutions were only able to fully automate simple coding cases[5]. CAC research to improve coding performances is thus still ongoing[6].

Furthermore, hospitals are increasingly digitizing and adopting national clinical terminology standards such as SNOMED CT in Australia. However, clinical terminologies are inadequate for serving the purposes of classification systems such as ICD-10 due to their immense size, granularity, complex hierarchies, and lack of reporting rules[7]. In order to realize the benefits of SNOMED CT, in particular for the purposes of clinical coding, SNOMED CT needs

[*] ICD-10-AM is based on the World Health Organisation ICD-10 system, as well as the Australian Classification of Health Interventions (ACHI), Australian Coding Standards (ACS) and ICD-O-3 (International Classification of Diseases for Oncology, 3rd edition).

to be linked and mapped to standard classification systems such as ICD-10. While maps can standardise the translation between coding systems, clinical coders still need to review and validate the ICD-10 codes to ensure accuracy with regard to the context of a specific patient encounter and compliance with coding guidelines[7].

This paper aims to bridge these gaps and challenges by investigating the effectiveness of a natural language processing (NLP)-based approach to CAC that specifically utilises mappings between SNOMED CT and ICD-10-AM. To this aim, hospital progress notes were used to provide the patient medical record for such an investigation. A broad spectrum of principal diagnosis codes (including complex and less prevalent clinical cases) with a reasonable spread across the ICD chapters[†] were selected for evaluation. The proposed approach achieved encouraging results (54.1% sensitivity; 70.2% positive predictive value) compared to what was projected as possible from a manual diagnosis code validation study (76.3% sensitivity). These results show the potential of NLP and SNOMED CT to ICD-10 maps for CAC.

Background

There has been a history of CAC technology and research. CAC solutions have included the use of software applications that facilitates and guides the coder to the "correct" code through to more complete coding solutions that extracts information in medical records to generate and suggest clinical codes[8-11]. NLP-based CAC methods generally involve a combination of rule-based and machine learning approaches.

Much research in clinical coding was from machine learning (and hybrid rule-based and machine learning) systems that automate ICD-9-CM[‡] code assignment from patient medical records[12-14]. These studies reported micro-F-measure coding performances of up to 89% from radiology reports, up to 54% from various physician authored documents, and 39.5% from discharge summaries. Other recent studies have also proposed deep learning methods for automated ICD-9-CM coding with competitive results[15,16]. The variations in reported performances were dependent on the context of the study such as density and number of unique codes considered, dataset and/or evaluation methods. As a result, systems across different studies were not comparable. In spite of this, CAC performances were generally low in an inpatient setting demonstrating the complexity of the task.

Other methods for improving code assignment have included the modification of confidence scores of codes by using propensity information from code co-occurrences[17], using PubMed to enrich the training data of infrequent diseases[18], and the use of structured information to complement the narrative patient record[16,19]. Incorporating structured data reflects clinical coding in practice where additional information, such as length of stay (LOS), history (e.g. episode/encounter number), procedures, demographics, was also used for assigning clinical codes. While augmenting structured data was important, micro-F-measure performances of around 40% using machine (and deep) learning models were achieved for assigning ICD-9-CM diagnostic codes[16,19]. Other CAC studies, beyond a hospital setting, have used death certificates to automatically code the ICD-10 cause-of-death[20-23].

Although numerous studies have targeted CAC in an in-patient setting using different datasets, ICD code systems and number of distinct codes, the performances of these systems were far below expert clinical coders. Most of these studies used machine learning but were not effective for diagnostic codes with limited training examples. The CAC method investigated in this paper used rule-based NLP approaches that do not involve the need for training data and at the same time allow for classifications of less prevalent diagnosis codes.

Furthermore, given the adoption of standardized clinical data using SNOMED CT in healthcare information systems, SNOMED CT to ICD mapping tables may be able to provide appropriate code set transformations useful for diagnosis coding. In fact, SNOMED CT to ICD-10 maps have been used in an emergency department setting to transform the clinically useful SNOMED CT code set in electronic medical records into an administratively appropriate ICD-10-AM code set for funding purposes[24]. Given the adoption of ICD-10-AM (Australian Modification) in Australian acute health facilities and the recent transition and adoption of ICD-10-CM (Clinical Modification) in the USA, such SNOMED CT to ICD-10 maps may also prove promising for diagnosis coding in an in-patient setting.

[†] ICD disease classification is hierarchical, with a small number of summary disease chapters. These ICD chapters were divided into a large number of more specific disease groupings (represented by 3-character codes). Most of the 3-character disease groupings can be further divided into an even larger number of very specific disease categories represented by 4-character and 5-character codes.

[‡] ICD-9-CM is the International Classification of Diseases, Ninth Revision, Clinical Modification

This study investigates a rule-based NLP method for normalizing the free-text medical records into SNOMED CT and then leverage the SNOMED CT to ICD-10 maps to automatically identify diagnosis codes. The extent of the gap between clinical coders and the proposed CAC approach for diagnostic coding was also investigated. To the best of our knowledge, there is no existing work that applies and evaluates an NLP-based, SNOMED CT to ICD-10 mapping approach for CAC in an in-patient setting.

Method

Data

Five years of patient encounter data (January 2011 – December 2015) was obtained from three Australian hospitals within the Gold Coast Hospital and Health Service (GCHHS)[§]. The dataset comprised 569,846 patient encounters and contained the following fields:

- age, gender, proxy unique identifier, ATSI (Aboriginal and Torres Strait Islander) status and associated patient demographics.

and admission details such as

- admission/discharge date and time, admission type, admission/discharge ward and unit, specialty, length of stay (LOS), principal and additional diagnosis codes (ICD-10-AM), procedure codes, and DRG.

The electronic medical record (EMR) data, which contained 8,644,797 progress notes, was provided from a separate system and required linking to the encounter data. As the EMR had a staged implementation across the three hospitals from November 2011 through to October 2012, not all patient encounters had associated progress notes.

The proxy unique identifier was consistent between the two datasets so linking was performed by matching the proxy unique identifier and the date ranges for the encounter. The linked set of progress notes to an encounter forms the data for a CAC system to process, analyse and suggest diagnostic codes for each of the patient encounters.

Ground truth

The principal and additional diagnostic codes, procedure codes and DRG codes were fields assigned by hospital clinical coders. In this study, the principal diagnostic codes formed the ground truth for evaluating the effectiveness of a CAC system. The diagnostic value of progress notes via a manual diagnosis code validation study was also conducted to place the performance of the system in context.

The ICD-10-AM diagnosis codes were either at a 3, 4 or 5-character level. The 3-character codes represent disease groupings while the 4 and 5-character codes represent very specific disease categories. While principal diagnosis codes were evaluated at their most specific level, we also truncated all the codes to their 3-character disease grouping level for analysis; for example, the code C34.1 (Malignant neoplasm: Upper lobe, bronchus or lung) would be truncated to C34 (Malignant neoplasm of bronchus and lung). Although this practice will result in 3-character codes that were regarded as invalid under the Australian Coding Standards, it will help in the analysis of system errors.

Diagnosis code validation

To better understand the data and determine the extent to which the ground truth principal diagnosis codes could be coded from progress notes, two in-house clinical coders (DT, MK), manually reviewed and coded a random subset of encounters pertaining to selected principal diagnosis codes. A broad selection of principal diagnosis codes was considered - see Table 1. The codes ranged from the more common through to the less frequent diagnosis codes across the range of ICD chapters forming a broad spectrum of diseases to investigate. Only diagnosis codes and encounters that were subjected to the full ICD coding process[**] were considered as candidate codes and encounters of interest. More specifically, the following codes/encounters were within the scope of the study:

[§] The data was obtained with research ethics approval from the GCHHS Human Research Ethics Committee.

[**] Full ICD coding process refers to encounters that have been coded by a professional clinical coder, taking into account all the available information within the medical record to assign codes based on the 'coding guidelines' or 'standards' for the assignment of principal diagnosis and additional diagnoses.

- Encounters with a care type of "Acute" or "Newborn" were in-scope as these care types were generally subjected to the full ICD coding process.
- Diagnoses where "batch" coding[††] was likely to be applied by the hospitals were excluded.

An initial review also prompted further limiting of encounters to those with a LOS of within a week[‡‡]. The LOS filtering still captured the majority of encounters but avoided those with a long LOS. Long LOS encounters generally contain a lot of observations and results with very limited information about confirmed diagnoses. The contents of these notes would likely contain lots of information about symptoms, conditions, history, social requirements, aged care placement/transfer/nursing home, and things to watch out for or to follow-up. The evaluation of this effect is of interest and would be the subject of future investigations.

Table 1. Principal diagnosis codes of interest.

ICD Chapter	Disease Classification	Code*
1	Certain infectious and parasitic diseases	A41.0 (14), A41.52 (6), A48.1 (5)
2	Neoplasms	C22.0 (50), C25.0 (45), C83.7 (11)
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism:	D50.0 (115)
4	Endocrine, nutritional and metabolic diseases:	E11.73 (141), E13.11 (3), E86 (310)
5	Mental and behavioural disorders	F05.1 (35), F10.7 (1), F31.1 (10), F31.3 (7)
6	Diseases of the nervous system	G04.8 (22), G10 (11), G31.2 (7), G37.3 (9), G45.9 (576)
9	Diseases of the circulatory system	I20.0 (599), I21.4 (1156), I50.0 (888), I60.3 (7)
10	Diseases of the respiratory system	J18.9 (1514), J22 (832)
11	Diseases of the digestive system	K80.10 (773), K92.2 (492)
12	Diseases of the skin and subcutaneous tissue	L03.13 (85), L89.3 (15)
13	Diseases of the musculoskeletal system and connective tissue	M86.94 (6)
16	Certain conditions originating in the perinatal period	P07.22 (3), P27.1 (4)
19	Injury, poisoning and certain other consequences of external causes	S72.03 (160), T42.4 (263)

* Numbers in brackets report the number of encounters after care type and LOS filtering.

Computer-assisted coding

Medtex, a medical text analytics platform that extracts and analyses key clinical information in medical text was adapted for the CAC task. Medtex has been used to support Cancer Registry tasks such as the cancer notification and the coding of notifications data from histopathology reports (including the ICD-O coding of topographies and morphologies)[25-27]. Medtex has also supported the ICD-10-AM coding of cause-of-deaths from death certificates[21,22].

Here, Medtex was adapted to provide a generic CAC service that derives ICD-10-AM diagnosis codes from narrative medical records for the diagnosis coding of in-patient hospitalization encounters. The approach used the core Medtex analysis engine[28], which utilises the Metamap software program[29] and the NegEx negation detection

[††] Batch coding refers to the coding of the same diagnosis code to all encounters from a particular hospital clinic due to the large quantity of similar cases.
[‡‡] The average LOS in the dataset for encounters with a care type of "Acute" or "Newborn" was 2.4 days.

algorithm[30] to standardize the free-text medical records by identifying UMLS[§§] and SNOMED CT concepts and their associated (positive or negative) assertions. Two cross-maps were used to translate the clinical concepts into ICD-10-AM codes. This was done via NLM's UMLS to ICD-10-AM mapping tables and an in-house version of SNOMED CT to ICD-10-AM maps[24] (curated and maintained over the years by CSIRO clinical terminologists).

A single hospital encounter would typically generate many SNOMED CT and UMLS concepts from its progress notes. As a result, multiple ICD-10-AM diagnosis codes may be generated for each patient encounter. In terms of CAC, these will form the code suggestions for validation by clinical coders. The proposed method would form a baseline system for assessing the feasibility and limitations of the approach. No further algorithm refinement or optimization was performed to tailor the approach for diagnosis coding.

Evaluation Measures

The effectiveness of the CAC approach was measured by evaluating the extent to which the ground truth principal diagnosis codes were being identified in the set of suggested codes. The effectiveness was measure using sensitivity (or recall) and positive predictive value (PPV or precision). A correct classification involved the principal diagnosis code being suggested by the CAC. Positive predictive value is the proportion of encounters with a system suggested principal diagnosis code that were correct, while sensitivity is the proportion of encounters with a given ground truth principal diagnosis code that were correctly suggested by the system. To provide a single, overall evaluation measure, precision and recall were combined into a third evaluation measure, F-measure.

For the CAC task at hand, sensitivity was considered to be more important as principal diagnosis codes were not missed by the system. False positive principal diagnosis codes, however, can be reviewed by clinical coders and be assigned as additional diagnosis codes or rejected if it was an inappropriate code suggestion.

Results and Discussion

Data analysis

The number of encounters increased 46.6% (98,103 to 143,779) over the 2011 to 2015 extract period. The increasing trend directly affects the clinical coding workload that is required for diagnosis coding.

There were 336,093 patient encounters with progress notes. A total of 4,977,092 progress notes were linked to these encounters, resulting in each encounter having on average 14.8 progress notes. The remaining 233,753 encounters without any associated progress notes were from periods when the EMR was yet to be implemented.

The progress notes comprised 1912 unique document types (e.g., 'INPT Clinical Note - Registered Nurse', 'ED Clinical Note - Resident')[***]. The majority of the progress notes were from health professionals such as nurses, dietitians, physiotherapists and social workers. Progress notes from Registrars, Consultants or Residents (or Interns) were also significant in numbers but were less in comparison to other health professionals.

Progress notes contained a lot of "noise": ungrammatical and fragmented free-text, shorthand notations, misspellings and punctuation errors. These challenges still remain an open clinical NLP research problem.

Ground truth analysis

Out of the 569,846 patient encounters over the five-year period, a total of 6675 unique principal diagnosis codes were recorded in the dataset. The frequency distribution of these codes shows that a small percentage of the principal diagnosis codes made up the majority of hospitalization encounters. In fact, the top 10 most prevalent principal diagnosis codes (0.15%) constituted approximately 30% of all encounters, the top 250 principal diagnosis codes (3.7%) made up approximately 70% of the diagnoses, and the top 500 (7.5%) made up approximately 80% of the encounters. The remaining 20% of encounters made up the long tail of other principal diagnosis codes.

Diagnosis code validation

Two in-house clinical coders, manually reviewed and coded a subset of encounters pertaining to 34 principal diagnosis codes of interest (Table 1). For each code, a small number of random encounters were reviewed to assess

[§§] Unified Medical Language System (UMLS) is a compendium of many controlled vocabularies in the biomedical sciences domain. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems. The UMLS is designed and maintained by the US National Library of Medicine (NLM).

[***] INPT: Inpatient, ED: Emergency Department

whether the ground truth principal diagnosis code could be reliably coded, solely from the documentation contained in an encounter's progress notes. The fraction of encounters reviewed that could be assigned the ground truth principal diagnosis code is presented in Table 2; The prevalence-weighted percentage for each ICD chapter takes into account the prevalence of the reviewed principal diagnosis codes of interest to provide a more representative measure of the utility of progress notes for diagnosis coding.

Table 2. Principal diagnosis code validation results.

ICD Chapter	Principal diagnosis	Principal diagnosis description	Fraction of encounters reviewed with Principal Diagnosis validated	Prevalence-weighted percentage of Principal Diagnoses validated from ICD chapter
1	A41.0	Sepsis due to Staphylococcus aureus	5/5	95.2%
	A41.52	Sepsis due to Pseudomonas	4/5	
	A48.1	Legionnaires' disease	5/5	
2	C22.0	Malignant neoplasm of liver and intrahepatic bile ducts	1/5	40.6%
	C25.0	Malignant neoplasm of pancreas	2/5	
	C83.7	Burkitt lymphoma	8/8	
3	D50.0	Iron deficiency anemia secondary to blood loss (chronic)	7/8	87.5%
4	E11.73	Type 2 diabetes mellitus with foot ulcer due to multiple causes	2/5 3/3*	54.1%
	E13.11	Other specified diabetes mellitus with ketoacidosis with coma	3/5	
	E86	Volume depletion		
5	F05.1	Delirium superimposed on dementia	3/5	43.0%
	F10.7	Mental and behavioural disorders due to use of alcohol	2/5*	
	F31.1	Bipolar disorder, current episode manic without psychotic features	0/5 1/5*	
	F31.3	Bipolar disorder, current episode depressed, mild or moderate severity		
6	G04.8	Other encephalitis, myelitis and encephalomyelitis	1/5*	95.1%
	G10	Huntington's disease	3/5	
	G31.2	Degeneration of nervous system due to alcohol	3/5	
	G37.3	Acute transverse myelitis in demyelinating disease of central nervous system	1/3*	
	G45.9	Transient cerebral ischemic attack, unspecified	5/5	
9	I20.0	Unstable angina	5/5	93.2%
	I21.4	Non-ST elevation (NSTEMI) myocardial infarction	5/5	
	I50.0	Heart failure	4/5	
	I60.3	Subarachnoid hemorrhage from posterior communicating artery	3/5	
10	J18.9	Pneumonia, unspecified organism	4/5	72.9%
	J22	Unspecified acute lower respiratory infection	3/5	
11	K80.10	Calculus of gallbladder with chronic cholecystitis without obstruction	2/5	47.8%
	K92.2	Gastrointestinal hemorrhage, unspecified	3/5	
12	L03.13	Cellulitis of lower limb	2/4	48.5%

	L89.3	Pressure ulcer of buttock	2/5	
13	M86.94	Unspecified osteomyelitis hand	3/5	60%
16	P07.22	Extreme immaturity of newborn, gestational age 23 completed weeks	2/3	57.1%
	P27.1	Chronic neonatal lung disease	2/4	
19	S72.03	Midcervical fracture of femur	1/1	87.6%
	T42.2	Poisoning by, adverse effect of and underdosing of succinimides and oxazolidinediones	4/5	

* May contain the review of encounters before care type and LOS filtering.

The results from the validation study showed that, on average, 63.4% (104 out of 164) of the encounters reviewed could be coded as the ground truth principal diagnosis from progress notes alone. When this result was prevalence weighted, approximately 76.3% of the encounters (equivalent to the sensitivity measure) could be coded as the ground truth principal diagnosis from progress notes alone. As such, there was approximately a 23.7% gap where other data sources may be able to complement and close this gap. There were also significant variations across the ICD Chapters indicating the variable usefulness of progress notes for the coding of particular diagnoses. Qualitative findings from the validation study are as follows:

- Effect of progress note type

The manual review identified that the most useful progress notes for clinical coding were written by the Registrar, Consultant or Resident (or Intern). Earlier entries of these notes in the record generally stated the known or admission diagnosis. Subsequent progress notes were unlikely to change, add or edit that original statement. Focusing on these progress note types would be most useful for finding candidate (principal) diagnoses.

Other progress note types were essentially written by nurses, dietitians, physiotherapists and social workers who do not make diagnostic decisions. These form the vast majority of the notes and often report on observations, which are of lower diagnostic value. These notes also serve as a communicative means between multiple care-givers, at handover, for follow-up, for completeness, and for safety and quality. This was especially found to be the case for encounters relating to "Mental and behavioural disorders" where substantial observations about things that have been or need to be done and things to be aware of regardless of the patient's diagnosis were being documented.

- Effect of 'single' progress note data source

Progress notes only represent one of many data sources from the patient's medical record. Diagnostic information may need to be derived from other data sources. This was particularly true for diagnoses relating to "Neoplasm" where information relevant to diagnosis coding was generally stored in other hospital information systems such as pathology and radiology information systems.

Furthermore, patient diagnosis for elective or planned admissions were generally already known (by the General Practitioner, specialist, ED, Outpatients, or Nursing Home) before hospital admission. These cases contain considerable assumed knowledge or information that existed outside of the progress notes for the patient. As a result, diagnoses information was unlikely to be found in the progress notes.

On the other hand, diagnoses that involve a test or measure (e.g., ECG, rate monitor) or have a care guideline or protocol (such as acute coronary syndrome or transient ischaemic attack (TIA)) were generally clearly stated in progress notes. As such, diagnoses relating to, for example, "Diseases of the circulatory system" were relatively easier to code from progress notes.

- Effect of clinical history

Many of the progress notes contained extensive patient history information (or family medical history) that was likely to contain 'old' diagnoses. This may influence the ability of a CAC system to discern the (principal) diagnosis for the current encounter. In spite of this, a number of reviewed encounters were coded to the historical condition, previously noted at an earlier time point. This should be done with caution as using clinical history will (i) inflate the count of diagnoses because it was already recorded and counted previously in earlier encounters and will be counted again, and (ii) disguise what had actually happened in the current encounter.

Computer-Assisted Coding

Table 3 presents the overall effectiveness of the CAC system for identifying the principal diagnosis code in the set of system suggested codes.

Table 3. Overall effectiveness for the CAC models.

Data	CAC using UMLS to ICD-10-AM map			CAC using SNOMED-CT to ICD-10-AM map			Manual diagnosis code validation of selected encounters**
	PPV	Sensitivity	F-measure	PPV	Sensitivity	F-measure	Sensitivity
All encounters	72.6%	36.5%	48.6%	69.6%	47.0%	56.1%	n/a
Care Type and LOS filtering*	72.8%	37.1%	49.2%	69.9%	47.9%	56.8%	76.3%
Care Type, LOS, Admit Status filtering*	73.1%	41.9%	53.3%	70.2%	54.1%	61.1%	n/a

* Encounter filters were as follows: Care Type (Acute, Newborn), LOS (≤7 days) and Admit Type (Emergency Admissions)
** Diagnosis coding validation result was analogous to the sensitivity statistic.

The results from the two maps show that the UMLS to ICD-10-AM map had slightly better PPV but lower sensitivity. The SNOMED CT to ICD-10-AM map, on the other hand, showed superior sensitivity, which was considered to be the more important metric as particular diagnoses were not missed by the CAC system. False positive principal diagnosis codes, which impacted on precision, can be reviewed by clinical coders and be assigned as additional diagnosis codes or rejected if it was an inappropriate code suggestion. These results were promising given that only 76.3% (prevalence weighted; equivalent to sensitivity) of encounters manually reviewed were able to be coded from progress notes. These results were from a NLP-based approach that was not fine-tuned (i.e. refined) or optimised for the coding task at hand. The fine-tuning of CAC systems has shown to improve coding accuracy over time[4]. Compared to other CAC studies that use machine learning approaches (see Background), the rule-based NLP approach using SNOMED CT to ICD-10-AM mapping tables has yielded encouraging results across a broad spectrum of diagnostic codes including complex and less prevalent codes.

The conditions where CAC can be successfully applied was dependent on whether the information about diagnoses were available in the medical record being processed. The improvement in evaluation results when considering certain encounter-level filters such as Care Type, LOS and Admit Status help removed encounters with irrelevant progress note contents; this reflected some of the findings from the manual diagnosis code validation study.

Other findings from the manual diagnosis code validation study were also reflected in the CAC results:

- Diagnoses that could be validated in progress notes also performed well by the CAC system (e.g., A48.1 “Legionnaires’ disease” with 80% sensitivity and 100% PPV, and C83.7 “Burkitt lymphoma” with 79% sensitivity and 100% PPV).
- For encounters where information was simply not present in the notes, the system failed to code them accordingly (e.g., F codes from “Mental and behavioural Disorder” with 0% sensitivity and PPV).

However, there were cases where the diagnosis information was in the notes but the system had difficulty in coding them (e.g., D50.0 “Iron deficiency anemia secondary to blood loss (chronic)” with 3% sensitivity and 100% PPV, and T42.4 “Poisoning by, adverse effect of and under-dosing of succinimides and oxazolidinediones” with 14% sensitivity and 100% PPV). These were possibly due to the diagnosis relation dependency on certain conditions or events, which may be better handled with machine learning models. Furthermore, if the CAC system could not understand a term or terms were spelt incorrectly, then the system may not have been able to assign the corresponding code. In contrast, there were also diagnoses where the CAC system did well while the selected samples for the review indicated insufficient information (e.g., C22.0 “Malignant neoplasm of liver and intrahepatic bile ducts” with 92% sensitivity and 100% PPV, and G37.3 “Acute transverse myelitis in demyelinating disease of central nervous system” with 82% sensitivity and 100% PPV).

To investigate the effect of the granularity of the principal diagnosis code output from the CAC system, Table 4 shows the effectiveness of the system for coding at a disease grouping level. Compared to Table 3, the marked

increase in sensitivity (approximately 10%) and PPV (approximately 3%) at a disease grouping level indicates that many of the diagnoses were actually classified at the disease grouping (3-character code) level correctly but not at the very specific disease category (4 and 5-character code) level. This is especially promising given the range of diagnosis codes across the ICD chapters investigated (including those with a low prevalence). This allows for targeted rule-base and/or machine learning algorithms to be developed to discriminate between these finer-grained categories to close the gap between the CAC system and clinical coders.

Table 4. Overall effectiveness for the rule-based models at a disease grouping (3-character code) level.

Data	CAC using UMLS to ICD-10-AM map			CAC using SNOMEDCT to ICD-10-AM map		
	PPV	Sensitivity	F-measure	PPV	Sensitivity	F-measure
All encounters	76.4%	46.0%	57.4%	73.2%	58.4%	65.0%
Care Type and LOS filtering*	76.4%	46.2%	57.6%	73.3%	58.9%	65.3%
Care Type, LOS, Admit Status filtering*	76.8%	51.9%	61.9%	73.7%	65.9%	69.6%

* Encounter filters were as follows: Care Type (Acute, Newborn), LOS (≤7 days) and Amit Type (Emergency Admissions)

Conclusion

Computer assisted coding (CAC) can improve the accuracy and productivity of clinical coding. The level of accuracy, especially for the more complex and/or less prevalent clinical coding cases across the full range of diagnosis codes, remains an open area of research. This research investigated this issue by assessing and analysing the performance of a NLP-based CAC coding approach that leveraged clinical terminologies under these settings.

The NLP-based approach used standard free-text to SNOMED CT normalisation methods and SNOMED CT to ICD-10 mapping tables for diagnostic coding. Promising results were obtained compared to a manual validation study. In contrast to other CAC studies that used machine learning approaches, the proposed rule-based NLP approach using SNOMED CT to ICD-10-AM mapping tables yielded encouraging results across a broad spectrum of diagnostic codes including complex and less prevalent codes. The approach was evaluated without any further fining tuning or optimisations of the method.

Future work will include (1) identification of additional data sources (including administrative data) to supplement the progress notes to enable the 23.7% gap in coding to be closed; and (2) improve current CAC performances to close the gap between the CAC system and clinical coders by investigating hybrid rule-based and machine learning models, addressing the noise issues in clinical notes, and incorporate the learnings from the diagnosis code validation study.

Acknowledgements

This research was done in partnership between the Australian e-Health Research Centre (AEHRC) at CSIRO and the Gold Coast Hospital and Health Service (GCHSS) within the Department of Health, Queensland Government. The authors acknowledge Michael Steine, Margaret Campbell, Kirsten Hinze, Adam Mende, Rachael Sewell, Nikki Roetman and Mark Tattam from GCHHS for their valuable advice, expert contributions and provision of data.

References

1. WA Clinical Coding Authority [Internet]. Department of Health, Government of Western Australia [Cited 14 February 2018]. Available from http://ww2.health.wa.gov.au/Articles/A_E/Clinical-Coding-Authority
2. Cheng P, Gilchrist A, Robinson KM, Paul L. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. Health Inf Manag J. 2009 Mar 1;38(1):35-46.
3. Australian Institute of Health and Welfare 2010. The coding workforce shortfall. Cat. no. HWL 46. Canberra.
4. Dougherty M, Seabold S, White SE. Study reveals hard facts on CAC. Journal of AHIMA. 2013 Jul;84(7):54-6.
5. Morsch M, Stoyla C, Landis M, Rogers S, Sheffer R, Vernon M, Jimmink M. Computer Assisted Coding At Its Limits - An Analysis of More Complex Coding Scenarios. Perspect Health Inf Manag, Computer Assisted Coding Conference Proceedings 2008

6. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc. 2010 Nov 1;17(6):646-51.
7. Bowman S. Coordination of SNOMED-CT and ICD-10: Getting the most out of electronic health record systems. *Journal of AHIMA* 76, no.7 (July-August 2005): 60-61.
8. 3M Codefinder Software [Internet]. 3M [Cited 14 February 2018]. Available from http://solutions.3m.com.au/wps/portal/3M/en_AU/HIS_AU/home/products-services/coding-grouping-reimbursement/codefinder/
9. 3M 360 Encompass System [Internet]. 3M [Cited 14 February 2018]. Available from http://www.3m.com/3M/en_US/360-encompass-system-us/computer-assisted-coding/
10. xPatterns [Internet]. Atigeo [Cited 14 February 2018]. Available from http://atigeo.com
11. Clintegrity CDI IT Solutions for ICD Conversion [Internet]. Nuance [Cited 14 February 2018]. Available from https://www.nuance.com/healthcare/clintegrity.html
12. Pestian JP, Brew C, Matykiewicz P, Hovermale DJ, Johnson N, Cohen KB, Duch W. A shared task involving multi-label classification of clinical free text. BioNLP 2007. pp. 97-104.
13. Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. J Am Med Inform Assoc. 2014 Mar 1;21(2):231-7
14. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif Intell Med. 2015 Oct 31;65(2):155-66.
15. Karimi S, Dai X, Hassanzadeh H, Nguyen A. Automatic Diagnosis Coding of Radiology Reports: A Comparison of Deep Learning and Conventional Classification Methods. BioNLP 2017. 2017:328-32.
16. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, Liu X, Sun M, Sundberg P, Yee H, Zhang K. Scalable and accurate deep learning for electronic health records. arXiv preprint arXiv:1801.07860. 2018 Jan 24.
17. Subotin M, Davis AR. A method for modeling co-occurrence propensity of clinical codes with application to ICD-10-PCS auto-coding. J Am Med Inform Assoc. 2016 Sep 1;23(5):866-71.
18. Zhang D, He D, Zhao S, Li L. Enhancing Automatic ICD-9-CM Code Assignment for Medical Texts with PubMed. BioNLP 2017. 2017:263-71.
19. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc. 2016 Apr 1;23(e1):e11-9.
20. Lavergne T, Névéol A, Robert A, Grouin C, Rey G, Zweigenbaum P. A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. BioTxtM 2016. 2016 Dec 11:60.
21. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 Classification of Cancers from Free-text Death Certificates. Int J Med Inform. (2015), 84(11):956–965. doi:10.1016/j.ijmedinf.2015.08.004
22. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, Truran D, Zhang M, Thackway S. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inform Decis Mak, 15:53, 2015
23. Duarte F, Martins B, Pinto CS, Silva MJ. Deep Neural Models for ICD-10 Coding of Death Certificates and Autopsy Reports in Free-Text. J Biomed Inform. 2018 Feb 26.
24. Lawley M, Truran D, Hansen D, Good N, Staibb A, Sullivanb C. SnoMAP: Pioneering the Path for Clinical Coding to Improve Patient Care. Studies in health technology and informatics. 2017 Aug 10;239:55-62.
25. Nguyen A, Moore J, O'Dwyer J, Philpot S. Assessing the Utility of Automatic Cancer Registry Notifications Data Extraction from Free-Text Pathology Reports. AMIA 2015 Annual Symposium, 2015
26. Nguyen A, Moore J, O'Dwyer J, Philpot S. Automated Cancer Registry Notifications: Validation of a Medical Text Analytics System for Identifying Patients with Cancer from a State-Wide Pathology Repository. AMIA 2016 Annual Symposium, 2016
27. Nguyen A, Lawley M, Hansen D, Bowman RV, Clarke BE, Duhig EE, Colquist S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc. 2010;17(4):440-445
28. Nguyen AN, Lawley MJ, Hansen DP, Colquist S. A simple pipeline application for identifying and negating SNOMED clinical terminology in free text. Health Informatics Conference, pp. 188-193, 2009.
29. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.
30. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301-10.