Generating SNOMED CT-encoded test data vignettes using Wikipedia

Edward Cheetham| NHS Digital (United Kingdom)

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS SPONSORED BY:
tpp

INTRO AND METHODS

END SHOW

TITLE SLIDE

Click on a bubble to navigate the e-poster

Intro and Methods

Methods

Results

Discussion

Abstract

This e-Poster demonstrates the way a structured knowledge resource (Wikipedia), SNOMED CT and a rudimentary NLP pipeline can be combined to generate simple coded clinical vignettes suitable for use as test data. The approach used, lessons learned, its strengths and weaknesses, opportunities for product improvement and further work are discussed.

Introduction

Good quality clinical test data is valuable for many health information technology development purposes. The characteristics that constitute ‘good quality’ clinical test data vary depending on the uses to which the data is put, but may include: being realistic/authentic, statistical representativeness, variety and versatility, availability in a usable form, and possibly include exceptional/unexpected features.

Clinical test data can be difficult to produce: anonymising real clinical data raises confidentiality concerns and generating suitable synthetic data by hand, whilst highly realistic, can be time-consuming and lack variety.

Wikipedia is an online encyclopaedia comprising more than 40 million articles on a wide range of topics. Although not a specialist healthcare resource, it has many entries describing human diseases, their features and management. Wikipedia content can be easily accessed several ways, including via an API.

This paper reports on using Wikipedia as an example of a structured clinical knowledge resource, analysing the text of sample entries using a simple natural language processing (NLP) pipeline with SNOMED CT as a structured entity recognition resource, and generating SNOMED CT coded clinical vignettes for use as test data.

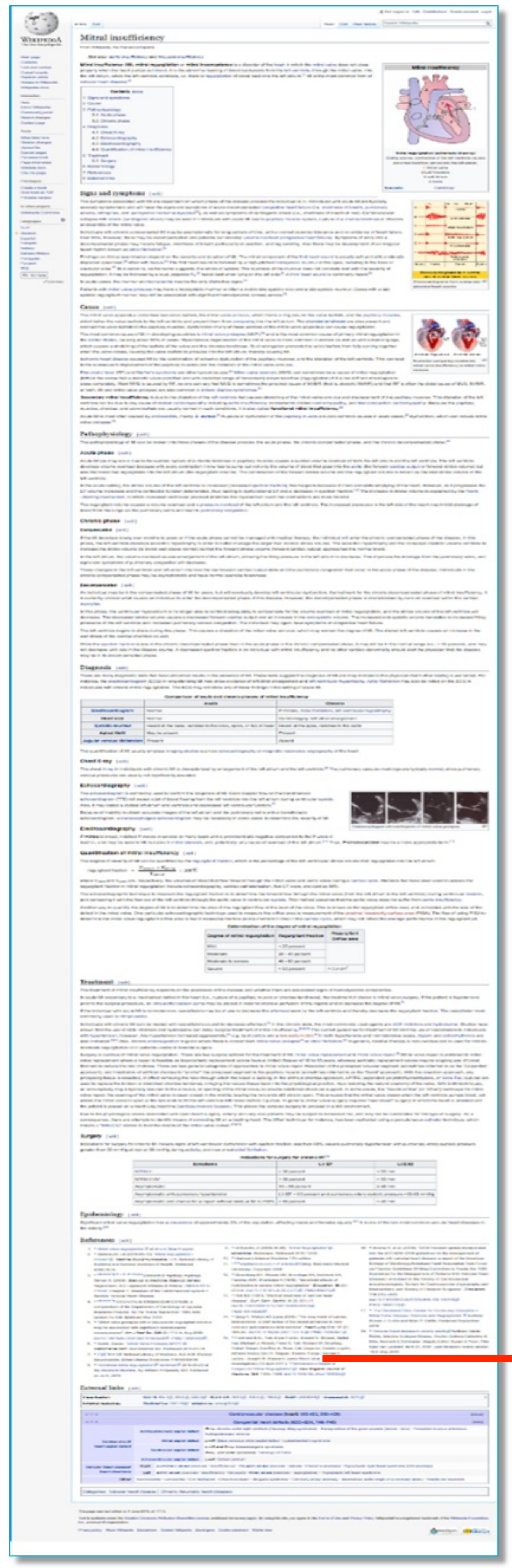
Methods

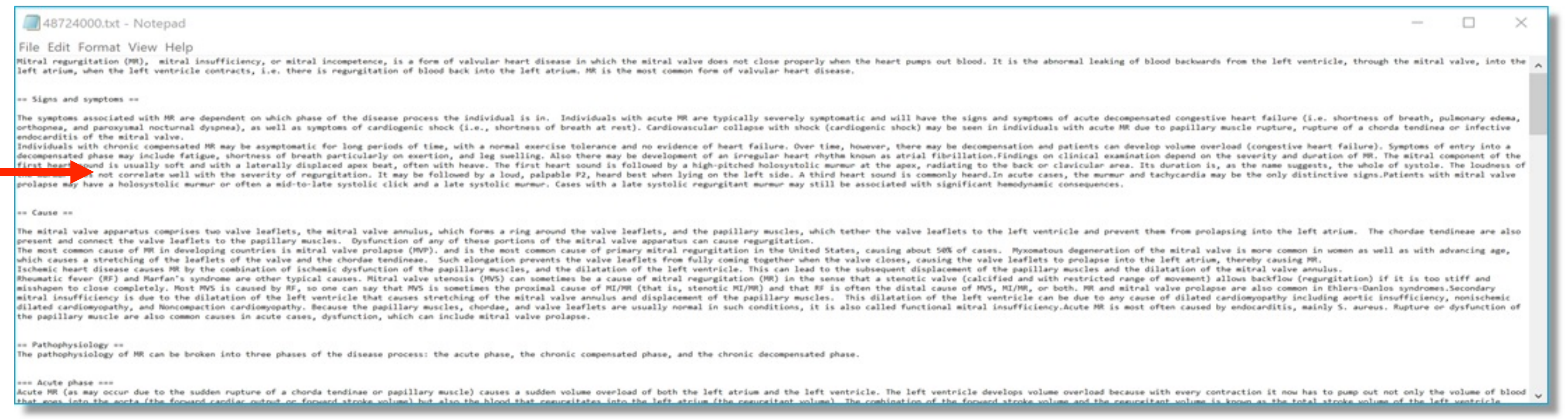
1. Wikipedia page detection

This project used the MediaWiki API [1] to automate the identification of Wikipedia entries that correspond to SNOMED CT-encoded entries.

A sample ‘search list’ to interrogate Wikipedia was generated from the membership of the SNOMED International GP/FP RefSet (4325 members). Terms for each concept were passed to the API search function, returning the plain text content of each matched page for further testing. In most cases this returned all the relevant page content, but occasionally bulleted list and tabular data is not returned with this approach.

Searches can match the submitted term exactly or can make use of Wikipedia’s auto-suggest feature. Generally, the auto-suggest feature increases sensitivity. Very occasionally matches are only identified with exact matching, so using both approaches and merging the results produces the largest yield.





```
Mitral regurgitation (MR),  mitral insufficiency, or mitral incompetence, is a form of valvular heart disease in which the mitral valve does not close properly when the heart pumps out blood. It is the abnormal leaking of blood backwards from the left ventricle, through the mitral valve, into the
left atrium, when the left ventricle contracts, i.e. there is regurgitation of blood back into the left atrium. MR is the most common form of valvular heart disease.


== Signs and symptoms ==

The symptoms associated with MR are dependent on which phase of the disease process the individual is in.  Individuals with acute MR are typically severely symptomatic and will have the signs and symptoms of acute decompensated congestive heart failure (i.e. shortness of breath, pulmonary edema,
orthopnea, and paroxysmal nocturnal dyspnea), as well as symptoms of cardiogenic shock (i.e., shortness of breath at rest). Cardiovascular collapse with shock (cardiogenic shock) may be seen in individuals with acute MR due to papillary muscle rupture, rupture of a chorda tendinea or infective
endocarditis of the mitral valve.
Individuals with chronic compensated MR may be asymptomatic for long periods of time, with a normal exercise tolerance and no evidence of heart failure. Over time, however, there may be decompensation and patients can develop volume overload (congestive heart failure). Symptoms of entry into a
decompensated phase may include fatigue, shortness of breath particularly on exertion, and leg swelling. Also there may be development of an irregular heart rhythm known as atrial fibrillation.Findings on clinical examination depend on the severity and duration of MR. The mitral component of the
first heart sound is usually soft and with a laterally displaced apex beat, often with heave. The first heart sound is followed by a high-pitched holosystolic murmur at the apex, radiating to the back or clavicular area. Its duration is, as the name suggests, the whole of systole. The loudness of
the murmur does not correlate well with the severity of regurgitation. It may be followed by a loud, palpable P2, heard best when lying on the left side. A third heart sound is commonly heard.In acute cases, the murmur and tachycardia may be the only distinctive signs.Patients with mitral valve
prolapse may have a holosystolic murmur or often a mid-to-late systolic click and a late systolic murmur. Cases with a late systolic regurgitant murmur may still be associated with significant hemodynamic consequences.


== Cause ==

The mitral valve apparatus comprises two valve leaflets, the mitral valve annulus, which forms a ring around the valve leaflets, and the papillary muscles, which tether the valve leaflets to the left ventricle and prevent them from prolapsing into the left atrium.  The chordae tendineae are also
present and connect the valve leaflets to the papillary muscles.  Dysfunction of any of these portions of the mitral valve apparatus can cause regurgitation.
The most common cause of MR in developing countries is mitral valve prolapse (MVP). and is the most common cause of primary mitral regurgitation in the United States, causing about 50% of cases.  Myxomatous degeneration of the mitral valve is more common in women as well as with advancing age,
which causes a stretching of the leaflets of the valve and the chordae tendineae.  Such elongation prevents the valve leaflets from fully coming together when the valve closes, causing the valve leaflets to prolapse into the left atrium, thereby causing MR.
Ischemic heart disease causes MR by the combination of ischemic dysfunction of the papillary muscles, and the dilatation of the left ventricle. This can lead to the subsequent displacement of the papillary muscles and the dilatation of the mitral valve annulus.
Rheumatic fever (RF) and Marfan's syndrome are other typical causes. Mitral valve stenosis (MVS) can sometimes be a cause of mitral regurgitation (MR) in the sense that a stenotic valve (calcified and with restricted range of movement) allows backflow (regurgitation) if it is too stiff and
misshapen to close completely. Most MVS is caused by RF, so one can say that MVS is sometimes the proximal cause of MI/MR (that is, stenotic MI/MR) and that RF is often the distal cause of MVS, MI/MR, or both. MR and mitral valve prolapse are also common in Ehlers-Danlos syndromes.Secondary
mitral insufficiency is due to the dilatation of the left ventricle that causes stretching of the mitral valve annulus and displacement of the papillary muscles.  This dilatation of the left ventricle can be due to any cause of dilated cardiomyopathy including aortic insufficiency, nonischemic
dilated cardiomyopathy, and Noncompaction cardiomyopathy. Because the papillary muscles, chordae, and valve leaflets are usually normal in such conditions, it is also called functional mitral insufficiency.Acute MR is most often caused by endocarditis, mainly S. aureus. Rupture or dysfunction of
the papillary muscle are also common causes in acute cases, dysfunction, which can include mitral valve prolapse.


== Pathophysiology ==
The pathophysiology of MR can be broken into three phases of the disease process: the acute phase, the chronic compensated phase, and the chronic decompensated phase.


=== Acute phase ===
Acute MR (as may occur due to the sudden rupture of a chorda tendinae or papillary muscle) causes a sudden volume overload of both the left atrium and the left ventricle. The left ventricle develops volume overload because with every contraction it now has to pump out not only the volume of blood
that goes into the aorta (the forward cardiac output or forward stroke volume) but also the blood that regurgitates into the left atrium (the regurgitant volume). The combination of the forward stroke volume and the regurgitant volume is known as the total stroke volume of the left ventricle
```

Plain text content from each matched Wikipedia entry (here Mitral Insufficiency [3]) can be retrieved using a suitable call to the MediaWiki API and stored locally for subsequent testing.

Generating SNOMED CT-encoded test data vignettes using Wikipedia

Edward Cheetham| NHS Digital (United Kingdom)

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS
SPONSORED BY:
tpp

METHODS

END SHOW

TITLE SLIDE

Click on a bubble to navigate the e-poster

Intro and Methods

Methods

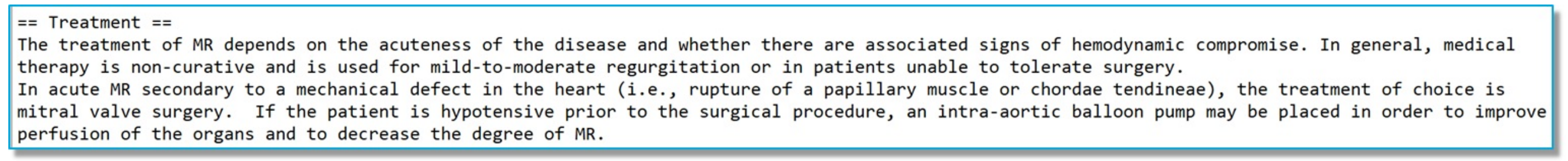
Results

Discussion

Methods

2. Page processing and identification of false positives matches

The majority of disease-based Wikipedia entries follow a predictable page structure. These sections and the contained content can be identified in the plain text returned, detected by the occurrence of the wiki syntax heading convention of two or more equal signs, for example:



```
== Treatment ==
The treatment of MR depends on the acuteness of the disease and whether there are associated signs of hemodynamic compromise. In general, medical
therapy is non-curative and is used for mild-to-moderate regurgitation or in patients unable to tolerate surgery.
In acute MR secondary to a mechanical defect in the heart (i.e., rupture of a papillary muscle or chordae tendineae), the treatment of choice is
mitral valve surgery.  If the patient is hypotensive prior to the surgical procedure, an intra-aortic balloon pump may be placed in order to improve
perfusion of the organs and to decrease the degree of MR.
```

The sections found in a true positive match are (* denotes mandatory):

- Signs and symptoms *
- Diagnosis
- Risk factors
- Causes
- Management or Treatment (this varies, so both can be tested for) *
- Complications

Failure to detect any of these sections is used indicate a false positive entry. For example, some disorders may match one of Wikipedia's 'List of ICD-9 codes' entries, and some disorders match non-health-related entries ('Burn of elbow' is matched to 'Professional wrestling attacks'). Identifying the sections is also important for subsequent processing.

3 NLP approach

3.1 Indexing of SNOMED CT

Using the Natural Language Toolkit [2], active acceptable descriptions of SNOMED CT (including en-US and en-GB variants) were indexed/normalised using the following steps:

- Stop word removal
- ASCII text conversion
- Word tokenisation
- Porter stemming
- Upper case normalisation
- Alphabetic key ordering

3.2 Wikipedia entry comparison

Corresponding normalisation steps were applied to the text from each Wikipedia entry section, along with sentence tokenisation, word equivalent application and the generation of 1-6 n-gram matching candidates. Preliminary experiments showed a significant drop in matching for any candidate strings longer than 6 tokens.

Part of speech (POS)-tagging was not used, nor were other sophisticated strategies such as context and negation detection. The matching approach took advantage of SNOMED CT semantic types: for example only looking for 'Finding', 'Disorder', 'Event' or 'Situation' concepts within 'Complication' sections.

Each n-gram generated was compared with the SNOMED CT index and any matches were stored.

3.3 Test data vignette preparation.

For each RefSet member for which a Wikipedia entry match was identified, test vignettes were prepared using the following approach:

1. As a minimum, each vignette required [1..*] presenting signs or symptoms, [1..1] 'diagnosis' (the name of the disease matched) and [1..*] treatments. This combination supports populating the essential/minimal features of a 'single episode summary', useful for producing coded data for testing discharge summary specifications.

2. Data cleansing and randomisation: Aspects of the NLP pipeline, in particular the use of structural stemming and not using POS-tagging, made the matching process vulnerable to false positives. Manual review of frequently-occurring matches in an initial sample of entries tested allowed many common false positives to be excluded from subsequent matching. For test data purposes the risks this exclusion approach introduced was felt to be acceptable. Any direct ancestors or descendants of the disorder concept used for matching were also removed. Optionally each vignette could be based on leaf concepts only (e.g. only using 'abdominal pain' where 'pain' has also been matched) accepting that this reduces the total number of matches available – and thus variety of vignettes. Depending on the total number of matches in each category, vignettes were then generated using a sample of available matches. If there was only one match then that single value would be used in every test vignette.

3. Where available, each vignette could then be augmented with other coded features, categorised according to their chapter/hierarchy locations within SNOMED CT (based on simple inclusion/exclusion constraints) and the entry section in which they were detected:

- Specific examination results
- Imaging, Microbiology or 'other' investigations
- Investigation results
- Associated conditions
- Causal organisms
- Risk factors
- Complications
- Treatments/interventions, further subdivided into surgical or non-surgical treatments
- Medicinal products (or very rarely substances where the product is not matched but a valid active ingredient can be). Specific 'infective vignettes' can be created where the causal organism is a bacterium and can be 'treated with' an antibacterial matched elsewhere in the entry.

Generating SNOMED CT-encoded test data vignettes using Wikipedia

Edward Cheetham| NHS Digital (United Kingdom)

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

E-POSTERS SPONSORED BY:
tpp

RESULTS

END SHOW

TITLE SLIDE

Click on a bubble to navigate the e-poster

Intro and Methods

Methods

Results

Discussion

Results

Quantitative results

Refset members and Wikipedia entries matched

Of the 4325 members of the July 2018 GP/FP RefSet (2751 disorders):

- 1588 (37% total, 57% disorders) match a Wikipedia page satisfying minimum vignette criteria.
- Matches were made to 923 unique Wikipedia entries
- 577 Wikipedia entries matched a unique RefSet member (by contrast, for example 10 distinct RefSet members are matched to the single Wikipedia entry for Chlamydia)

SNOMED CT matches to text of Wikipedia entries.

The number of codes matched within each section and filtered by SCT chapter/category (minus manually detected exceptions) are shown in the following table:

Wikipedia section	SCT Chapters matched	Total matches	Unique matches	Sample of most frequent matches
Signs and symptoms	Clinical finding	25181	2369	Pain, Swelling, Fever
Diagnosis	Clinical finding, evaluation procedure, diagnostic procedure	25488	3100	Imaging, plain radiography, MRI, pain, ultrasonography, blood test, microbial culture
Risk factors	Clinical finding, disorder, substance, procedure, situation, event	9359	1962	Alcoholism, problem drinker, alcohol (substance), smoking, obesity
Causes	Disorder, drug, event, organism, situation, substance	61573	7987	Trauma, gene, alcohol, obesity
Management or Treatment	Drug, procedure, situation	50259	7213	Procedure [injection, exercises, drainage, surgical repair] Drug [steroid, NSAID, vitamin]
Complication	finding, disorder event, situation	17286	2895	Bleeding, pain, death, injury, swelling

Qualitative evaluation

In terms of clinical plausibility, synthetic 'clinical vignettes' produced this way can reasonably be criticised – some of the coded content does not withstand even casual clinical scrutiny and each 'story' implicitly compresses what may be years of care into a short episode. However, even such a 'fiction' could still provide SNOMED CT-encoded sample data suitable for many (in particular technical) test purposes.

As an example, and based on analysis of the Wikipedia page for Mitral valve regurgitation [3], the following sample coded vignette can be generated (codes are not shown to ease readability):

'Test record' section or subsection	SCT terms matched to Wikipedia text (sampled from larger set of matches for this entry)
Signs and symptoms	Oedema, Regurgitation, Paroxysmal nocturnal dyspnoea, Heart irregular, Tachycardia, Retching and Orthopnoea
Specific examination results	Pansystolic murmur, Third heart sound and Late systolic murmur
Associated conditions (merged from 'causes' and 'complications')	Congestive heart failure, Atrial fibrillation and Rupture of chordae tendineae
Imaging investigations	Magnetic resonance imaging (MRI) of vessels, Transoesophageal echocardiography, Colour doppler ultrasound and Angiocardiography
Other investigations	Electrocardiographic monitoring
Medication	Hydralazine, Diuretic and Angiotensin-converting enzyme inhibitor agent
Non-surgical interventions	Stretching and Perfusion
Surgical interventions	Ring annuloplasty, Construction of shunt and Repair of mitral valve

For many entries it is possible to match multiple SCT terms against Wikipedia text. Where this is the case, greater variety of vignette can be achieved by sampling random values against each section or subsection. As such, multiple similar but varied vignettes can be generated for Mitral valve regurgitation.

Generating SNOMED CT-encoded test data vignettes using Wikipedia

Edward Cheetham| NHS Digital (United Kingdom)

SNOMED CT Expo 2019
Kuala Lumpur | Oct 31-Nov 1

Click on a bubble to navigate the e-poster

Intro and Methods

Methods

Results

Discussion

Discussion

General observations

The vignettes produced so far are clearly far from perfect. They do however demonstrate that even a trivial NLP pipeline can be used to identify and extract a rich variety of SNOMED CT codable content from structured textual resources.

Referring back to the desirable features of test data listed in the introduction, the approach used in this paper appears to meet some features better than others. Experience of producing vignettes suggests that (tested against the 'discharge summary' ambition):

- 'Variety and versatility' are fairly well satisfied – it is certainly easy to generate *large numbers* of different vignettes.
- Availability in a usable form – effectively the goal of the exercise. The intention is to generate structured and coded vignettes, and this is achieved.
- Statistical representativeness – unlikely. This feature would require additional data indicating disease frequencies, age groups etc.
- Realistic/authentic – cannot be guaranteed and depends on the vignette generated. If the use case requires authenticity then it may well still be more efficient to generate vignettes automatically and inspected by experts rather than have experts hand-craft each vignette.
- Include exceptional/unexpected features – these may be present, but cannot be introduced 'to order'.

Opportunities for improvement

Performance could clearly be improved by a number of strategies, including:

- The application of even modestly more sophisticated NLP techniques could help improve match yield and fidelity (with, for example, greater consideration of context and negation), and improved word form and sentence analysis is likely to improve matching accuracy.
- By returning matches against other SNOMED CT content, such as detecting anatomy, morphology and actions, it would be possible to experiment with generating compositional matches (although this would also be dependent on more sophisticated NLP).
- Use of alternative, more specialist clinical knowledge sources may provide better substrate for analysis (recognising that this work concentrated on Wikipedia because of its ease of access.
- As indicated above, addition of metadata such as disease frequencies and associated patient age group data could be used to improve authenticity and statistical representativeness.

Further opportunities and application

The declared goal of this work was to investigate the production of clinical data vignettes for inclusion in the testing of communication standards during their development. The richer and more varied the payload used in testing a clinical communication specification, the better the testing process. Hand-written test data is time-consuming to produce, and may find itself being reused and therefore systematically fail to detect certain error types. Whilst undertaking the work it has become apparent that other opportunities, based on the routine interrogation of textual knowledge resources using SNOMED CT, exist. These include:

- Putting the 'clinical' back into clinical terminology:
  - SNOMED CT's way of representing and organising its content is valuable – distinguishing content into variably broad semantic types would not have been possible without it, but these organising principles do not 'speak' to all users.
  - SNOMED CT's ideal is stable compartmentalisation of ontologically distinct ideas. The ability to generate 'clinical stories' beyond the immediate neighbourhood of a disease animates SNOMED CT's content, illustrating its relevance to the dynamics of normal clinical activity.
- Improving content development processes
  - The work presented uses and tests features of the SNOMED CT data, which will in turn generate exceptions and errors: some of these errors will be because the process is wrong, but some will be because the data is wrong, and such errors can be reported back to SNOMED CT developers for correction.
  - By shining a light on relatively neglected parts of the terminology, it is possible that long-standing undetected errors in content will be revealed.
- It may also be possible to generate code sets that complement those already being developed.
  - If a clinical group develops a list of diagnoses relevant to their practice, currently that is all they will receive.
  - It may be possible to use a similar process to that described above, based on a seed of diagnostic codes of interest, to produce candidate sets of corresponding treatments, symptoms, causes and complications, thus simplifying, accelerating and enhancing current development processes.

References

[1] https://www.mediawiki.org/wiki/API:Main_page + https://pypi.org/project/wikipedia/
[2] https://www.nltk.org/
[3] https://en.wikipedia.org/wiki/Mitral_insufficiency