Masters of Research – Year 2 Project

Characterisation of surgeries beyond billing codes

Faculty of Medicine and Health Sciences
October 9th 2015

Author:
Georgina Kennedy
mq43913202

Supervisor:
Dr. Blanca Gallego Luxan[1]

Main Text: 19242 words
Abstract: 248 words
Number of Figures: 10

Keywords: Surgical Procedures, Ontological Modelling, Natural Language Processing, Patient Safety, Automated Surveillance

[1]Centre for Health Informatics, *Australian Institute of Health Innovation*, Macquarie University, Sydney, NSW, Australia

Conflict of Interest Statement

The author declares that there are no conflicts of interest.

Statement of Candidate

This thesis is presented as a partial fulfilment to the requirements for the degree *Masters of Research*.

I certify that the work in this thesis entitled *Characterisation of surgeries beyond billing codes* has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that this thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself has been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

Georgina Kennedy (43913202)

9th October 2015

Abstract

The record of what occurred during a surgical procedure is typically represented in the electronic health record as a combination of unstructured text blocks (the operative report) with limited associated structured data. Billing codes fail to account for significant variance in procedures, thus although much of this information is valuable for real-time patient safety interventions, it is infrequently available for automated analysis.

The selection of an appropriate ontological model provides a good foundation for effective information extraction and knowledge representation, allowing high quality inference and knowledge based concept identification. Through gap analysis and statistical analysis of the content of a corpus of operative notes, SNOMED CT has been selected as the most appropriate knowledge model for automated information extraction in this domain.

To successfully apply statistical natural language processing (NLP) methods developed on one corpus to another type of text, one must assume that there is a sufficient degree of similarity between the texts, both syntactically and semantically. From this, a determination is drawn as to the applicability of existing clinical NLP tools to the operative report. General clinical text was found to be not representative of the writing observed in operative reports.

From this theoretical foundation, text classifiers were developed to demonstrate the feasibility of automatically encoding a subset of SNOMED CT terms in operative reports. Classification performance was high for detection of surgical specialty and open or closed procedures (f-score 0.965, 0.931 respectively); however, the detection of laterality was more reliable through heuristic methods.

Acknowledgement

De-identified clinical records used in this research were provided by:

1. Macquarie University Hospital
2. The i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organised by Dr. Özlem Uzuner, i2b2 and SUNY.

Ethics approval for research

Ethics application reference number 5201500351 was reviewed by the Macquarie University Human Research Ethics Committee (HREC (Medical Sciences)) at its meeting on the 23rd of April, 2015. This research was found to meet the requirements set out in the *National Statement on Ethical Conduct in Human Research* (2007 – Updated March 2014) (the *National Statement*) and approval was provided on the 28th of May, 2015.

Contents

List of Figures

List of Tables

Definitions and Abbreviations

ANZASM	Australian and New Zealand Audit of Surgical Mortality
AE	Adverse Event: An injury related to medical management, in contrast to complications of disease. Medical management includes all aspects of care, including diagnosis and treatment, failure to diagnose or treat, and the systems and equipment used to deliver care. Adverse events may be preventable or non-preventable [1].
CCAM	Classification Commune des Actes Médicaux
CEC	Clinical Excellence Commission
CHASM	Collaborating Hospitals' Audit of Surgical Mortality
CPT	Current Procedural Terminology (USA)
DRG	Diagnosis-Related Group
EHR	Electronic Health Record
GASP	GALEN model for Surgical Procedures
GALEN	Generalized Architecture for Languages, Encyclopedias and Nomenclatures
GRAIL	GALEN Representation and Integration Language
HIS	Health Information System: any electronic system within the clinical setting that is used for data generation, compilation, analysis and synthesis, and communication and use [2]. This includes but is not limited to electronic health records, incident management systems, anaesthesia information management systems, clinical decision support systems, electronic ordering and prescription systems, billing and administrative systems, laboratory systems and centralised purpose-specific repositories such as disease or mortality registries.

ICD	International Classification of Diseases
ICPS	International Classification for Patient Safety
IHTSDO	International Health Terminology Standards Development Organisation
IIMS	Incident Information Management System
KL	Kullback Leibler Divergence
MBS	Medicare Benefit Schedule (Australia)
MUH	Macquarie University Hospital
MUHOH	Macquarie University Hospital Operative Notes corpus
NLP	Natural Language Processing
NSW	New South Wales (Australia)
OR	Operating Room
OWL	Web Ontology Language
Patient Safety Intervention	Any strategy that is implemented in order to improve patient safety – examples include clinician education, policy implementation guideline development, feedback of outcomes to healthcare providers, audits and accountability measures.
Patient Safety System	A subset of health information systems that are used for the purpose of surveillance, prediction or improvement of patient safety outcomes.
RACS	Royal Australasian College of Surgeons
RCT	Randomised Controlled Trials
SNOMED CT	Systematized Nomenclature Of MEDicine – Clinical Terms

Surgical Notes / Operative Reports	Reports written or dictated by surgeons at the completion of a surgical procedure. These may be unstructured or semi-structured text blocks and typically contain details of the procedure performed, observations and outcomes.
Surgical Procedure	Procedures that are used for diagnosis or treatment that involve incision, puncture, entry into a body cavity [3]. Excluding non-surgical interventional procedures, which use ionising, electromagnetic or acoustic energy only.
SSI	Surgical Site Infections
SVM	Support Vector Machine
UMLS	Unified Medical Language System
WHO	World Health Organization

1 Introduction

1.1 Background

Worldwide, it has been estimated that approximately 1 in 25 people undergo a surgical procedure annually [4]. Estimates vary as to the rate of major surgical complications, but even the most conservative range of 3-5% (for developed and developing nations respectively) represents a significant public health concern [4].

In order to be able to reduce rates of all kinds of surgical complications, it is first necessary to fully understand and be able to detect these events and the context in which they took place, if only to have a meaningful baseline against which to measure effectiveness of interventions. These monitoring and surveillance activities are inherently rooted in the type and quality of data that is captured to describe procedures performed and the related patient characteristics.

When compared to the practice of pharmacovigilance and medical device monitoring, the surveillance of patient safety within surgical procedures is relatively immature. This is partially the result of legislation and oversight, which is required to control the competing financial, intellectual property and patient safety interests of pharmaceutical product manufacturers, and is therefore well established. There is limited parallel incentive in the development of surgical procedures, and as a consequence safety monitoring tends to be at the institution level rather than globally or federally mandated and coordinated.

There have been recent efforts by the World Health Organization (WHO) to remediate this issue and initiate a coordinated, harmonised effort for adverse event characterisation (for both surgical and non-surgical events) through the International Classification for Patient Safety (ICPS) [5]. This endeavour was initiated at the 55$^{\text{th}}$ World Health Assembly held in 2002; however, at the date of writing ICPS remains an incomplete classification, and as such, has yet to be taken up by any member states [6].

In New South Wales (NSW), Australia, deaths that occur whilst a patient is under the responsibility of a surgeon (in either the public or private healthcare system) must be reported to the Collaborating Hospitals' Audit of Surgical Mortality (CHASM)[2] program as a requirement of the continuing professional development program of the Royal Australasian College of Surgeons (RACS). This is then collated at a regional level by the RACS Australian and New Zealand Audit of Surgical Mortality (ANZASM) program.

Adverse events (AEs) (surgical or otherwise) not leading to death, occurring in the NSW public hospital system, are reported to a state-wide Incident Information Management System (IIMS), however there does not exist similar centralised all-encompassing AE reporting system for the private healthcare system. Section 20L of the Health Administration Act 1982 [7] defines a subset of incidents that must have a root cause analysis report submitted directly to the Ministry of Health, however this comprises only the most severe and egregious events (such as wrong patient/wrong site surgery). Many serious AEs, surgical or otherwise, are thereby subject to very limited external monitoring practices, as seen in Figure 1 [7, 8, 9].

[2]Under the joint remit of the Clinical Excellence Commission (CEC) and Royal Australasian College of Surgeons.

Additionally, drug-related adverse events tend to be simpler to characterise. Drug formulation, dosage, frequency, route of administration, cumulative dosage and concomitant medications, in concert with general patient characteristics, are likely to be sufficient details to fully describe at least the causative factors of a given event [10, 11]. It is noteworthy that all of these elements are discrete variables, which makes them attractive for centralised data comparison, automated processing and signal analysis. This allows even rare events to be confidently detected and causality established by centralised data analyses, such as the international reporting systems established at the Uppsala Monitoring Centre [12].

In contrast, the reality of surgical procedures is nuanced and variable, and this is only partially captured by the categorical nature of modern clinical coding systems. In practice, the same coded event may represent a variety of techniques, tools, approaches, treatment rationales and/or patient risk categories. See, for example, MBS item code 38498 for a coronary artery bypass, which may be performed *either via a median sternotomy or other minimally invasive technique*. This quite explicitly does not stipulate a specific surgical approach, tools used, number and type of incisions etc. Each of these elements may be a contributory factor in any adverse outcomes experienced.

Clinical systems are by their nature not well suited to being restricted to entirely structured data – both the format and content of diagnostic judgements and clinical narrative are diverse and unpredictable. The imposition of limitations on the ability of clinical personnel to create unstructured narrative text will necessarily degrade the scope of data captured by these systems. This will have a significant impact on the quality and completeness of record keeping, to the point that it may interfere with patient safety, as future diagnostic and therapeutic decisions will be based on an incomplete picture.

The unavoidable existence of clinically significant information in an unstructured format has lead to the design of many surgical patient safety interventions and monitoring systems resorting to expert manual record review. This is an expensive and time-consuming undertaking, with results not available in anything close to real-time [13]. Refer, for example, to the CHASM Surgical Case Form [14] that provides up to two whole pages to describe the course to patient death, which is then submitted for as many as three rounds of manual peer-review.

It is possible under certain circumstances to build patient safety interventions based on structured data alone, which has obvious benefits for automated analysis and therefore cost and timeliness of the overall system. This is particularly effective for the most commonly occurring adverse outcomes such as surgical site infections (SSI), where tools built using only coded data may even outperform manual review [15]. Avoiding the use of unstructured data entirely, however, is not suited to all cases, as the diagnostic and anatomical rationale for decisions is typically unavailable within structured data, which limits the generalisability of results.

The primary goal of this project is thus to analyse the current practice of surgical characterisation in a tertiary care private hospital in Sydney, Australia. This will be compared

[3]Note that the CHASM and IIMS procedures are wholly independent, and that "Surgical Deaths" and "Adverse Events" do not represent mutually exclusive categories – thus an incident may fall into one, other or both categories, potentially triggering multiple reports.

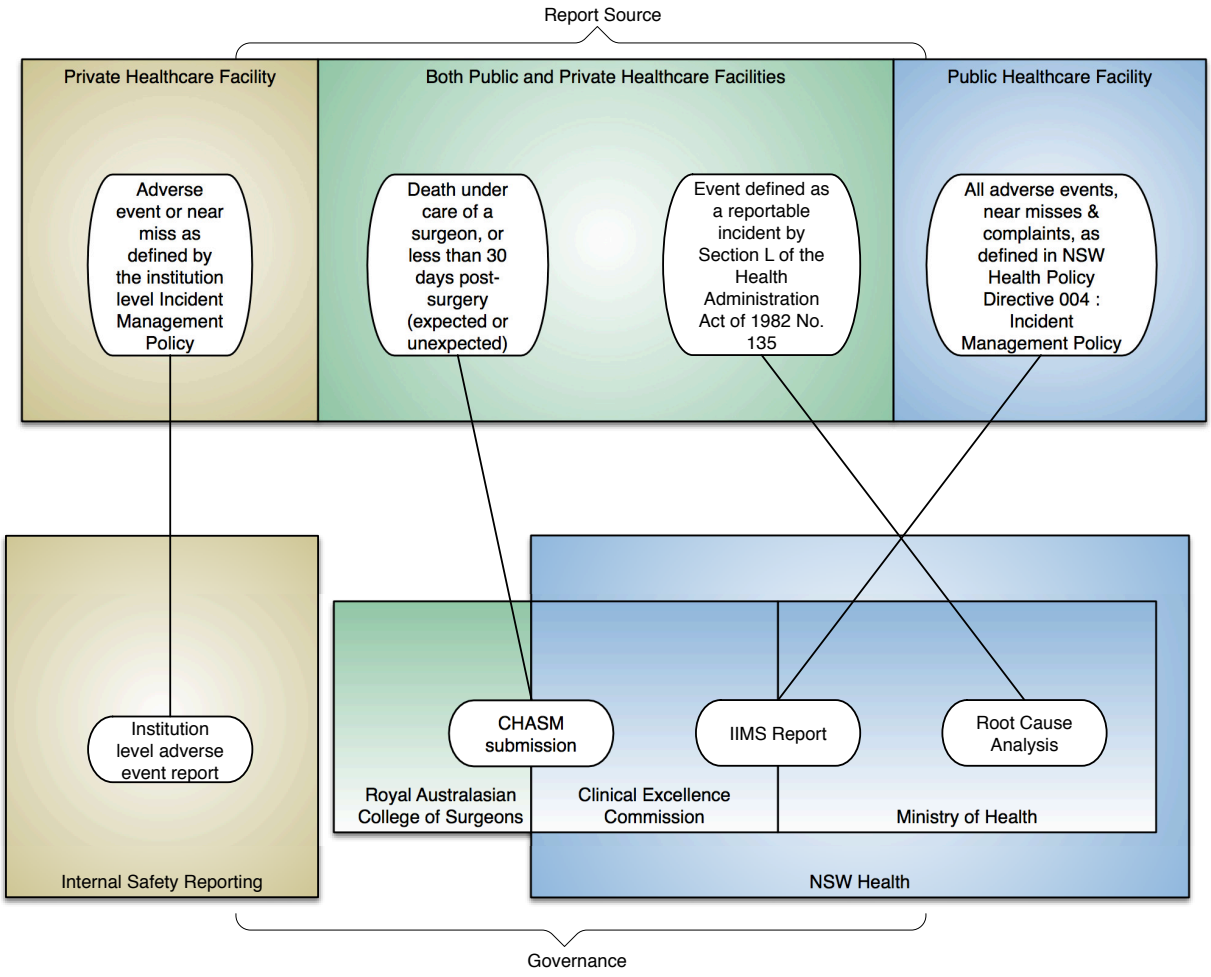


Figure 1: Reporting of adverse events and surgical deaths in NSW

to the state of the art, and strategies will be proposed to develop a more satisfactory practice that will facilitate the development of effective, generalisable patient safety and monitoring systems.

1.2 Problem statement

There is a gap in the current practice for representation of surgical procedures, where characterisation by free text, together with billing codes (as is typical within EHR systems) is insufficient for many types of automated clinical inference.

This ineffective representation is particularly impactful in the monitoring of patient safety within surgical procedures. This is a highly manual task, which is mired in inefficiencies such as double data entry and manual notifications and reporting. There are significant efficiency gains that can be made by the automation of this process.

Some part of the data required for automation of patient safety monitoring is present in surgical notes and theoretically can be detected by the application of natural language processing, however this is not commonly or systematically implemented at the time of writing. To facilitate meaningful analysis, a target structure must be carefully chosen to capture knowledge from the text in an orderly manner. This target structure must allow statistical epidemiological comparison both between procedures, and at the sub-procedural level.

In addition, there are many concerns regarding the systematic evaluation of the efficacy and safety of new surgical techniques. Randomised controlled trials (RCTs) in surgical innovation are the exception rather than the rule. Particularly for incremental changes, it is likely that the evaluation will be extremely limited and be primarily comprised of uncontrolled, low-quality data in the form of case studies [16, 17, 18]. Setting aside the larger issues of designing acceptable RCT protocols for surgical procedures, and the reluctance of surgeons and patients to accept randomisation, the standardisation of surgical representation will streamline the comparison of complex techniques. Automation of this task will also greatly enrich the available data for retrospective observational studies.

It is therefore valuable to explore existing ontological models as they apply to surgical notes and identify the best fit for extracting knowledge. This is expected to reap rewards in the form of improved accuracy of text processing.

In addition to improving patient safety monitoring practices, mapping free text notes to a structured data format allows for:

- Access to the notes in a database form, which speeds up processing and simplifies analysis.
- More robust de-identification – as free-text notes have limitations for protection of patient identity, thereby increasing availability for research.
- Availability of additional data for statistical research – this is especially relevant in studies of comparative effectiveness of surgical procedures, as researchers have historically struggled to compare incremental changes and per-clinician preferences in surgical technique. By unlocking details of what actually occurred during a surgical procedure

at a far higher granularity than what is currently available, this will provide a clearer basis for empirical guideline development.

- Information retrieval – patient charts may contain many hundreds of reports, and it is not possible for clinicians to synthesise information in a timely fashion without some assistance via automated categorisation.
- Machine learning for the purpose of predicting patient outcomes.
- Automated monitoring and surveillance of concepts which are typically not available as part of the coded data.
- Quantification of variance between procedures at the clinician or group level for the dual purposes of detection of fraudulent claims and improving quality and adherence to guidelines.

1.3 Research gap

There are many ontological models that have been proposed to represent surgical procedures for different purposes; however, there is no clear best candidate model to capture the detail from operative reports that is required for monitoring and safety procedures.

A robust analysis of potential target ontological models that are in common usage within clinical systems is required. From this, the most promising candidate target structure(s) for the purpose of characterising surgeries can be identified and modifications proposed as necessary.

Likewise, although natural language processing (NLP) techniques have been applied with success to the clinical domain for a number of purposes, there is limited available work that has dealt specifically with surgical notes.

1.4 Aim

The overall objective of this research is to perform a thorough review of current encoding systems and NLP in the surgical domain. This will be done in order to identify limitations in currently available methods and tools for the purpose of performing automated encoding of surgical notes. From this basis a strategy to achieve full automation will be proposed, prioritisation of development activities and preliminary algorithm design will be undertaken and evaluated as a proof of concept.

When selecting the appropriate target model, priority shall be given to tasks related to patient safety and monitoring; however, the ideal model will be flexible and reusable for many purposes.

2 Characterisation of surgical procedures

Before making any judgement of the adequacy of surgical procedure characterisation within the Macquarie University Hospital (MUH) surgical department, an understanding of what is typical, achievable and desirable is required. This section therefore details current methodologies used to represent surgical procedures in clinical information systems. The first part of the chapter describes the type of data that is collected, methods of collection and purposes for which the data is accessed at MUH. This is followed by an analysis of the types of representations commonly used both in Australia and globally, and then concludes with selection of the desired method of knowledge representation.

2.1 The surgical data environment at Macquarie University Hospital

The capacity of a given knowledge model to provide valuable structure to the information present in a system is highly dependent upon the ability of the designers to anticipate the tasks for which the knowledge will be accessed. A good understanding of the eventual targets of these data allows the selection of a flexible, extensible template-oriented knowledge model, which closely aligns with both user goals and the data present in the system [19].

Although the goals of this project are focused in the safety and quality domain, it must be accepted that there are many diverse users within the hospital (and beyond) who will benefit from accessing this surgical knowledge. Allowing for their needs where possible will increase the likelihood of successful uptake of any proposed changes. Taking the view of the hospital as a closed system for the purpose of simplification, surgical data collection and downstream data use cases are described below.

2.1.1 Surgical data: collection

A simplified overview of the peri-operative surgical data capture procedure at MUH is provided in Figure 2. This process was documented during the observation of two surgical procedures at MUH in March 2015, with follow-up interviews with key personnel. It does not include ongoing post-operative data that is captured as the patient's recovery is monitored, nor does it include data that is captured pre-admission or in the admission clinic. The only formal encoding system used to characterise procedures that is used during the peri-operative phase is the MBS – all other data is stored as either semi-structured text blocks, or as discrete data points within system-specific data models.

Under the current system, the majority of peri-operative data capture activities are the responsibility of the scout (non-sterile) nurse and anaesthetist. The scout nurse is responsible for the capture of pre-operative checklists, counts, operation timings, post-operative handover and nursing care instructions. Checklists will include data such as the provision of informed consent, confirmation of correct patient, intended surgical site and other required

preparatory checkpoints. Counts include not only the precautionary pre/post counts required for the prevention of retained objects, but also inventory management for sterile tools, implantable objects and consumable items. These latter counts are generally performed by barcode scanning within the MUH systems. Operation timings include times entering and leaving theatre, start and end times of the operation and other specialty-specific times such as tourniquet start and end time.

The anaesthetist will capture details of administration of medications while the patient is in theatre, plus details of transfusions, monitoring methods, fluid requirements and lines placed. The systems used by the anaesthetist also monitor the patient's vital signs, although this information is not always stored and therefore may not be available within the patient record for later reference, unless particular snapshots are taken and saved by the anaesthetist.

In a typical surgical procedure, the surgeon (or assisting surgeon) does not enter any information into the EHR system until the main portion of the procedure has concluded. At this point, they will either dictate or transcribe detailed notes, which describe what has occurred and been observed during the surgical procedure. These notes are captured under the following synoptic headings: *Operation Performed*, *Details of Operation*, *Operation Findings*, *Closure*, *Particulars of Tubes/Drains/Catheters Left Insitu.*, *Wound Classification*, *Post-operative Instructions Surgeon.* The surgeon will also make a classification of the procedure as a set of MBS codes that they believe accurately represent the surgery as it occurred. This forms the primary block of free-text known as the Operative Report, which is the object of interest for this project.

2.1.2 Surgical data: use cases

The following use cases (Figure 3) were developed based on informal interviews with staff members from MUH, including surgeons, nurses, IT professionals, accounting team members, quality assurance managers and medical coders. They represent the most common day-to-day activities within the hospital.

Use cases regarding the booking and administration of the operating rooms, consumable items, implants etc. are omitted here, as they are upstream of the creation of the surgical report. This diagram instead details those tasks that create surgical data (gold), refer to the surgical data as a primary source (blue) or secondary/coded source (purple). Research activities are also excluded from this analysis, as MUH does not currently have a standardised procedure for research data requests at this time and instead deals with each research project in an ad-hoc fashion.

Downstream of the primary surgical data capture process, secondary encoding of the procedure is performed by a team of dedicated medical coders who derive appropriate procedure codes from the International Classification of Diseases 10^{th} Revision (ICD-10) on a per-admission basis. The coders rely on the entirety of the clinical record when making a judgement on which ICD-10 codes to apply. From the ICD-10 codes, diagnosis-related groups (DRGs) are generated. These DRGs are the primary source of reimbursement from health funds to the hospital. In the instance that a given DRG is outside of the negotiated agreement between the hospital and private health fund, the MBS codes will then be implemented as a

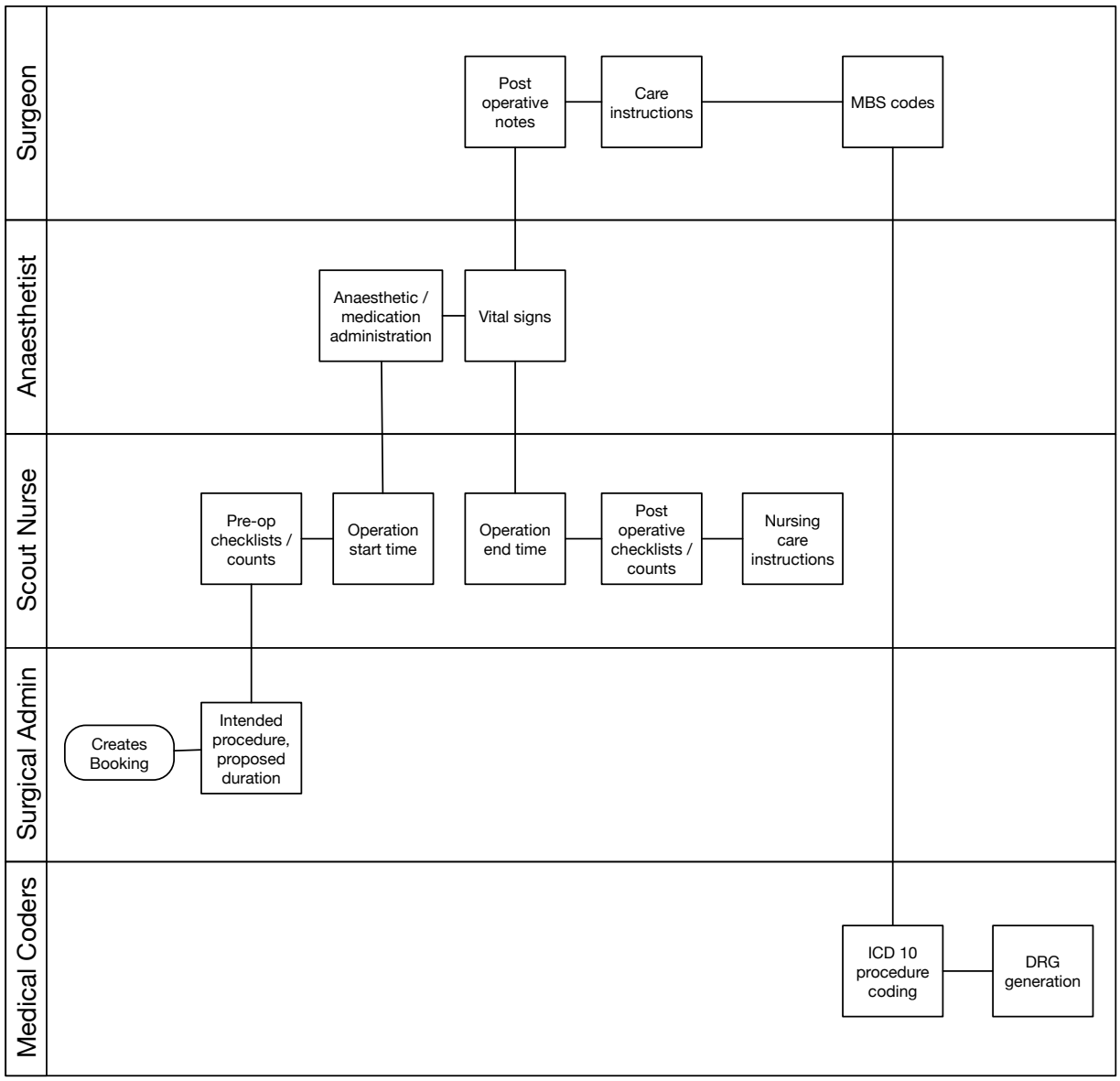


Figure 2: MUH surgical data capture process

substitute billing method. In MUH, the surgeons bill for their services separately to the bill that is generated by the hospital accounting department, as is typical in private hospitals in Australia. The surgeon's bill is based entirely on MBS procedure codes defined as described above.

In the context of creating a more effective characterisation of surgical procedures, it is important to understand all use cases that access the surgical data products and their eventual downstream targets such that the needs of all users can be met. It is clear from the use case analysis that two primary challenges are presented to the users of this data – first, that clinical processes are highly manual, with many users referring back to blocks of text and transcribing between systems, which requires re-familiarisation and interpretation for each activity; and secondly, that many administrative processes rely on only on secondary, manually created data products (i.e. MBS, ICD-10 codes, DRGs), which exposes these activities to the risk of inadequate granularity of characterisation and potentially reduces the validity of these analyses. ICD-10 codes and DRGs are also only available post-discharge, as the billing activities are undertaken on a per admission basis. In the area of patient safety, both the aggregate quality and audit activities and investigations of specific clinical incidents would benefit from having timely, automated access to reliable structured data upon which to base this analysis.

2.2 Knowledge representation

It is the hypothesis of this project that a better characterisation exists, or can be developed, which both increases the efficiency of clinical processes and improves the accuracy and timeliness of inferences made by the administrative processes. This requires a survey and evaluation of available knowledge representation methods, which will be presented here.

2.2.1 What is an ontology?

The development of ontologies can be seen as an effort to provide structure to knowledge by formalising the definition of concepts, their relationships and behaviours. In this instance we take a broad interpretation of the generally accepted definition of what constitutes an ontology, which is *a formal, explicit specification of a shared conceptualisation* [20]. This allows the inclusion of both the simplest code sets providing only a semantic mapping (concept to associated code), to fully abstracted models of a domain that derive not only elements and their definition but also their attributes, relations, restrictions and axioms, such that complex logical inferences can be made.

In some contexts, the classification of a coding systems as ontologies is rejected as they do not meet the *open world assumption* [21], and often do not contain a minimal hierarchical relationship definition [22]. For the purposes of this project, however, it is necessary to include the wider definition in order to adequately capture the status quo.

Ontologies typically fall along a spectrum of characterisations from the general to the specific, depending on the type and granularity of information they contain (Figure 4). All ontologies reviewed for this project are specific to the clinical domain and thus domain ontologies. Domain ontologies built upon a widely implemented upper level ontology should

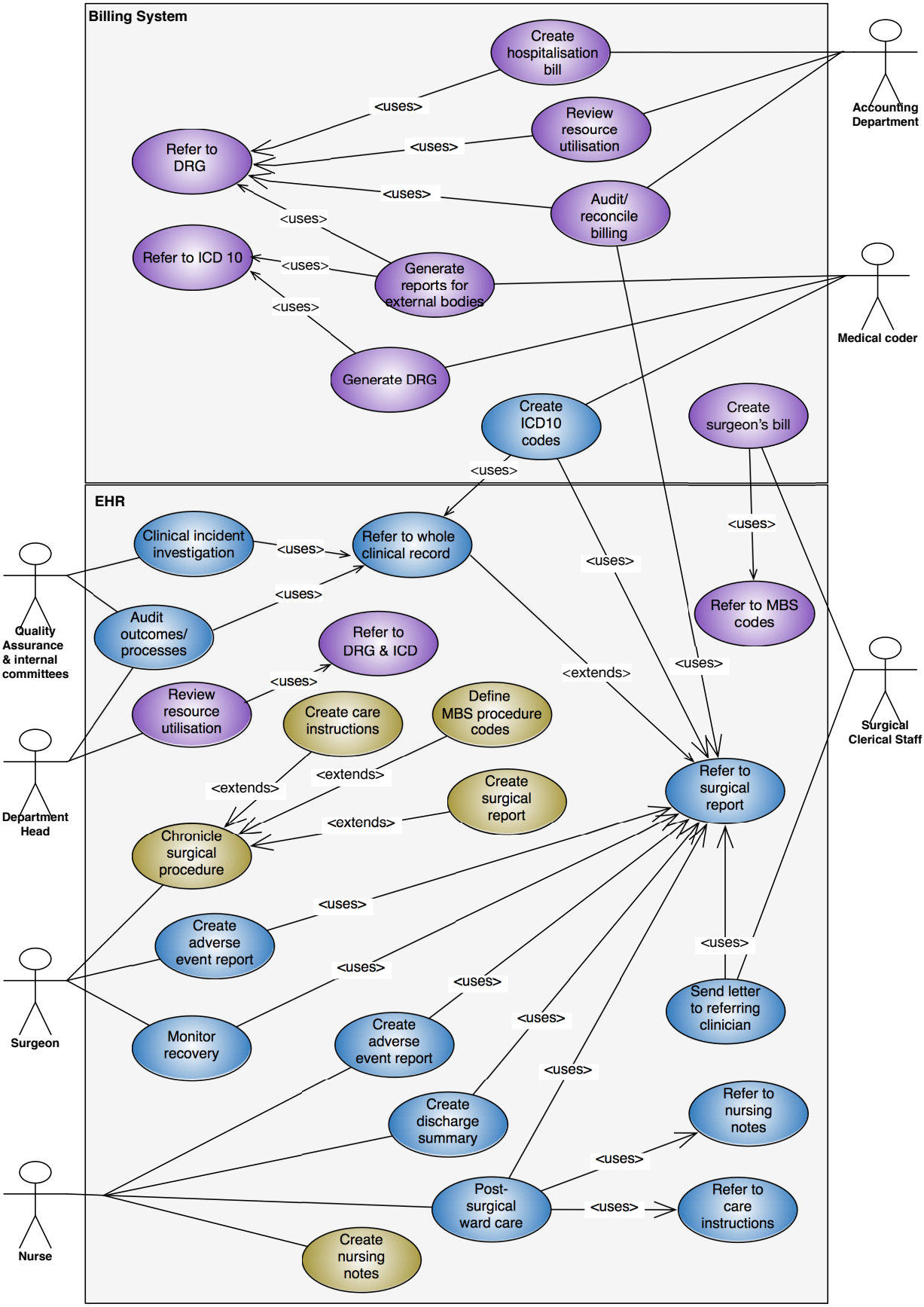


Figure 3: Surgical data use cases

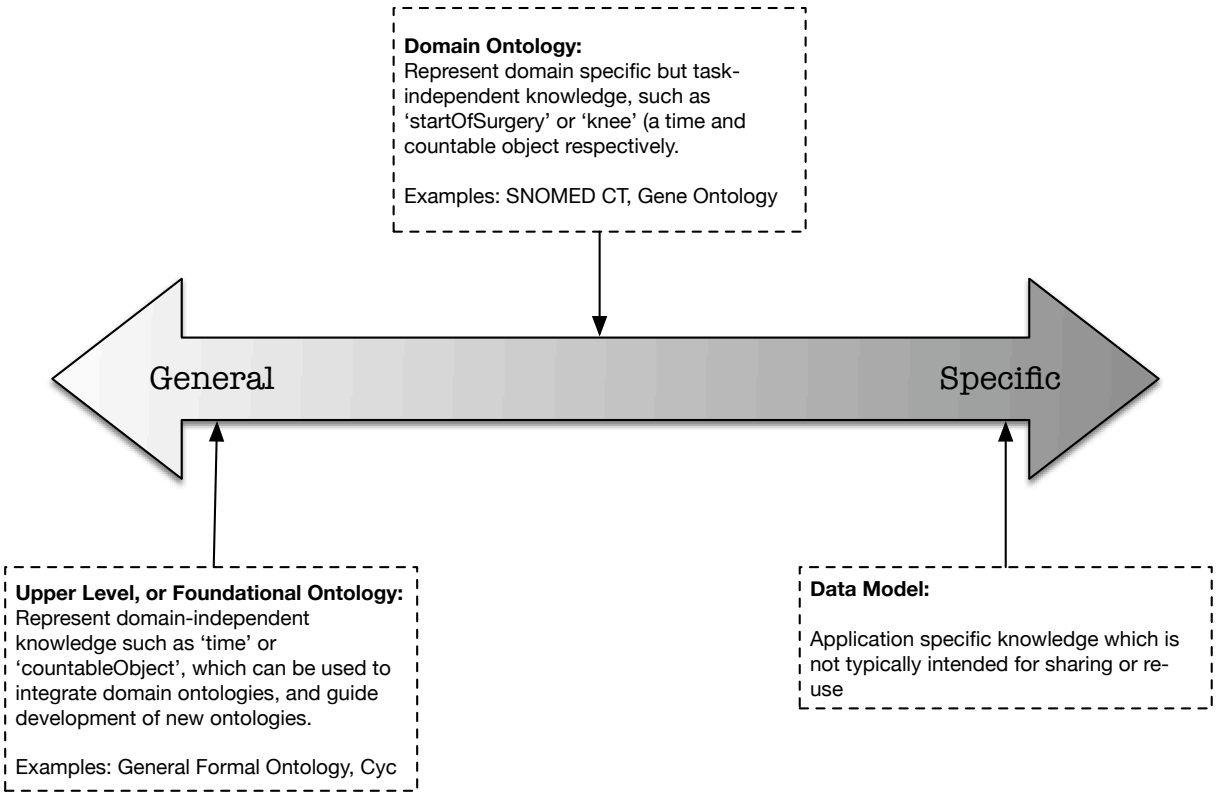


Figure 4: Generalisability of ontologies

be viewed preferentially, given the subsequent capacity for integration with other systems and uniformity of development. Data models are rejected out of hand, as they are unsuitable for the sharing of data outside of the context in which it was collected and thus contribute to the siloing of clinical data, which is contrary to the goals of this project.

2.2.2 Evaluation of ontologies

A trade-off must be made between the comprehensiveness of a domain ontology, allowing effective and meaningful intersystem communication, and its simplicity, allowing greater generalisability [23]. Ontology designers seek to achieve maximal coverage with the smallest possible spanning set of concept definitions with their associated attributes and relationships, for the closest possible approximation of the target domain. This is known as the minimal ontological commitment.

The evaluation of ontologies presents a significant challenge, in particular due to the fact that the goal of knowledge formalisation is typically sharing and interoperability; by definition, ontology designers frequently do not know the full set of use cases in all possible source and target systems in advance. This makes it difficult to create test scenarios that provide coverage of all situations. Efforts have been made to formulate a methodology for ontology evaluation that can achieve a high level of confidence in the quality and suitability for a given purpose of a specified ontology.

Evaluation is required for both the content and structure of an ontology, in order to

ensure that it is compliant to the modelled world [24]. The criteria described below have been compiled from the reviews [25, 26, 27, 28], and together represent a thorough methodology for systematic evaluation. Note that it may not be possible to apply all of these criteria to every situation, and the criteria must be balanced against both what is practical given the scope of the target ontology, and the tasks for which the ontology is designed and used.

Table 2: Evaluation of ontologies

Criteria	Description
Completeness, Richness and Granularity	For a given use case, is the definition of the modelled domain adequate? This may not require 100% coverage if (i) use cases are sufficiently well known or (ii) the ontology can be post-coordinated to provide additional definitions that are not present in the base ontology.
Conciseness	Are there redundant definitions in the ontology (explicit or inferred)? Are irrelevant definitions included? Is there a minimal ontological commitment – i.e. does the ontology make as few claims as possible about the domain in order to communicate knowledge unambiguously?
Consistency, Accuracy and Coherence	Is it possible to create definitions or inferences that are contradictory (internal or logical consistency)? Is the conceptualisation consistent with the real world (accuracy, or external consistency)?
Expandability and Adaptability	How much effort is required to add new definitions, or expand existing definitions, without requiring adjustment of properties that already exist?
Clarity	Are the definitions unambiguous and independent of context? Is the ontology intelligible to a human user with domain relevant expertise?
Computational Efficiency	Is the ontology designed in such a way that logical inferences are efficient? Are measures of semantic similarity and distance possible to calculate in a meaningful fashion? How quickly can queries be processed?

Organisational Fitness	Does the ontology align with other ontologies that are already in use in the organisation? This 'goodness of fit' may be in terms of compatibility with the same upper level ontology, capacity for cross mapping if there is any domain overlap, or potentially the same ontology expression language. Is it easy to access by all potential stakeholders? This includes a consideration of closed vs open source and how freely accessible the licensing requirements are.

It is not feasible for this project to apply this full evaluation methodology to all ontologies in common use for surgical representation. This is due to the prohibitive scope of the content and detailed review required, and the fact that a number of the questions require implementation in a real-world application in use to fully assess. This chapter will instead apply these concepts at a high level in order to immediately rule out the majority of candidate ontologies for obvious deficiencies. A more detailed evaluation is applied only to those models most likely to prove effective.

Regardless of the scope of evaluation and testing applied in the ontology development phase, it is still likely that for any non-trivial ontology errors will be present. Some types of errors can be automatically detected [29] and should not be present in any validated, published ontology. Importantly, circularity errors (where a class can be defined as a subclass or superclass element of itself), grammatical redundancy errors (where a semantically equivalent definition can be formed in multiple ways), and partition errors (where an element has been defined inconsistently with respect to a parent class – such the same concept being as defined in 2 disjoint partitions) should be detectable prior to release [30].

Other errors are a matter of inconsistent definition, semantic redundancy or omission [31], and thus can only be detected when compared to other domain definitions – either the knowledge of experts, or some other predefined knowledge base. This includes artifacts such as incompleteness or semantic inconsistency errors. These errors are more likely to make it into a production release, and there should be a well-established process for their resolution.

In addition to these measures of quality, there is of course also the measure of fitness for purpose. It is not effective to select even the most optimal ontology if the representation does not provide the tools required to meet the current goal.

2.2.3 Benefits of ontologies in a clinical context

Formal knowledge models are desirable in a surgical context for a number of different purposes such as automation, simulation or clinical decision support. For all of the models described in this chapter, a key concept is the idea of making elements of the surgical procedure observable and understandable by computers. At some level, each of these models provides a shared understanding of the surgical domain that can be unambiguously communicated between computers and people [20].

A well-designed ontological model is also highly valuable in creating a representation that is both system and language independent by providing semantic interoperability [32]. This allows users and applications from different institutions and/or countries to pool and

exchange information seamlessly. In the patient safety domain, and especially with respect to rare events, increasing the size of available data sets allows researchers to accurately ascribe causality of adverse events, and to discern predictive patterns with higher confidence.

Ontologies have also been found to provide context and meaning to NLP applications, as it becomes possible to move from the purely statistical and syntactical models into a model that understands at some level the concepts to which it is being applied [33]. Taxonomic NLP moves beyond a statistical lexicon and instead leverages resources containing hierarchical semantic associations. The primary work in this field has been associated with the Semantic Web.

An understanding of the ideal representation of the target knowledge as provided by a well-designed ontology is also useful in the action of segmenting and categorising reports at predefined, meaningful levels. This is an important step in an end-to-end NLP processing pipeline.

2.3 Clinical ontologies in common usage

This section provides a description of the subdomain ontology types that are in common use in the clinical setting, with examples. For the most part, biomedical ontologies such as the Gene Ontology [34] are expressly excluded – although there may be some conceptual overlap, their design for non-clinical purposes limits their applicability.

2.3.1 Administrative and operational

For the purpose of billing, the representation of procedures is relatively mature, as a limited number of dimensions are needed to adequately represent an element or event. In fact, many simple clinical interactions may be represented by a single code, for example, take what is likely to be the most general of all MBS codes – *3: consultation at consulting rooms*. Whilst this is sufficient for the purposes of administration, it gives only the barest picture of what occurred and does not represent any clinical observations or inferences. This is obviously inadequate granularity for the purpose of monitoring patient safety.

Surgical procedures will often paint a somewhat more detailed picture, as a number of codes are listed to capture a more complex clinical encounter. For example, a laparoscopic total colectomy may require the MBS codes *32090*, *30393*, *18262* and *32012* to encode the progression of the operation. There are still many salient clinical features which are left uncharacterised in this instance, however – take, for example, the size, number and nature of any polyps observed. It is outside of the scope of billing codes, however it may be relevant when analysing the underlying features of an adverse event.

Similarly in the quality domain, billing codes are insufficient to understand the intent and progression of a surgery. Criteria such as average surgical duration, resource utilisation or appropriateness of selected treatment are often applied in order to judge the performance of the surgical team. A uni-dimensional basis for this type of analysis such as billing codes oversimplifies the nature of the surgical process and disincentivises clinicians from accepting complex cases. This is often tempered with a risk adjustment algorithm, taking into

account other patient factors that influence outcomes and approaches; however, there is a significant amount of information available that is still being left out of these analyses in most cases.

The coverage or completeness of representation provided by billing codes is not the only issue with their use in characterising complex surgical events. In addition, the hierarchical structure and granularity of a given ontological system is a key factor in determining the nature of the systems and inference engines that can be successfully built on its basis. Systems designed purely for billing purposes tend to be relatively flat structures, where the categorisation lacks foundational generic concepts.

Within the MBS, and its US counterpart the CPT, for example, procedures belong to categories, groups and subgroups; however, there are no taxonomic relationships or predicative rules on the basis of which any clinical inference could be reasonably made. The ICD system, portions of which are prevalent as a billing system in a large part of the world, was originally designed as a diagnostic classification standard for clinical and research purposes and therefore does not suffer from this same limitation.

The DRGs, which are the basis for the hospital level billing system at MUH, are by definition related to diagnosis rather than procedure. This may be useful in a risk-adjustment algorithm, however has limited bearing when representing the activities undertaken during admission in detail – the level of granularity is intentionally low in order to avoid cost-based reimbursement and instead require hospitals to assume some of the uncertainty of resource allocation in the process of healthcare provision.

Billing codes should also be acceptable to characterise procedures for the purposes of order entry – in combination with clinician preference card systems and inventory/ward management they are sufficient for operational purposes.

2.3.2 Clinical decision support

This is the development of knowledge-based systems that implement an inference engine in order to provide decision support. These systems define the clinical reality in terms of:

1. A question to be answered.
2. The synthesis of data relevant to this question.
3. Rationale applied to the data in order to formulate an answer [35].

A well-designed ontology may be able to encode all of these elements in one, or there may be an individual representation of each component.

The key differentiator between ontologies underpinning clinical decision support and those designed for other purposes is the requirement to model clinical practice guidelines as a reference standard. This must be done in a form that can be automatically compared with the current patient state in order to formulate decisions.

A review of clinical decision support models found that most models studied could be simplified to the following primitives: actions, decisions, patient states and execution states, with decisions and actions forming the critical requirements [36]. These systems are likely to have a positive impact on patient safety, as surgical procedure variability is reduced and practices brought in line with the best available knowledge.

A model based on these primitives is a closer approximation to the type of data that is captured in surgical reports than billing and administration models; however, there is no real consensus on the best model upon which to build clinical decision support applications [37]. This heterogeneity may be the result of the additional difficulty experienced when integrating a model deeply into the clinical workflow, rather than as an endpoint. This lack of clear direction means that it is somewhat high risk to base the goal representation on a clinical decision support model due to the unpredictability of which models (if any) will have continued support.

2.3.3 Adverse event reporting

The World Health Organization (WHO) has published draft guidelines for the reporting of adverse events. The goal of this document is to provide a broadly accepted basis for steps and systems required to allow effective reporting of adverse events beyond the institute level [1]. This guideline does not prescribe a universal model of representation of adverse events (and the processes of care leading up to them) for the purpose of reporting – advocating a mix of structured data and free text as appropriate to each particular situation. It does, however, list a number of classification systems, which can be applied to facilitate post-hoc analysis.

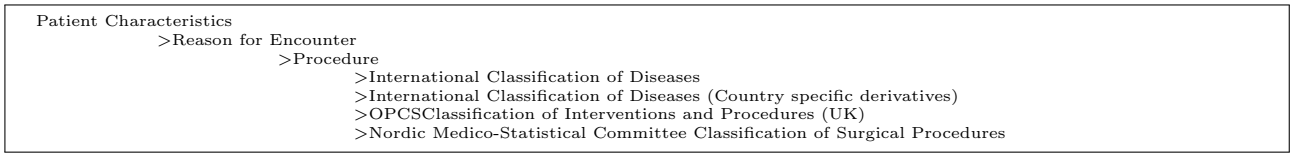


Figure 5: Surgical characterisation sub-hierarchy of ICPS

A more recent project by the WHO, in collaboration with the European Union (EU) is the International Classification for Patient Safety (ICPS) [5]. The purpose of this project is to design a model for the representation of all adverse events within national and international reporting systems. Surgical procedures form only a small part of the representation provided by this adverse event model. See Figure 5 for the sub-hierarchy of surgical representation within the ICPS model. In this model, if the 'reason for encounter' for a given event is a 'procedure', the relevant national coding system (ICD-10-AM for the purposes of an Australian provider) is leveraged directly.

This model is not suitable for the purposes of this project, as it requires significant additional data not available directly from the operative report, and conversely does not have the capacity to represent details of the procedures beyond the ICD (or similar) coding systems. It does, however, implement techniques that provide a good basis for the definition of a model of surgical procedures, namely an OWL2 representation and natural language generation functionality for faithful translations. It is understood that at the point of writing,

the ICPS development track has been put on indefinite hiatus due to lack of funding [38]. Despite this fact, compatibility with the ICPS model is still a desirable feature in any selected ontology due to both its rigorous design and the fact that it is the model closest to a general acceptance at an international level.

Current legislation in NSW requires manual reporting of surgical mortality to CHASM, within the Clinical Excellence Commission – itself a part of NSW Health. This includes not only adverse events, but also deaths caused by natural disease progression whilst the patient is under the care of a surgeon, or up to 30 days post surgery. The CHASM reporting process is initiated by either the local health district (in the public health system) or an individual hospital (for private institutions).

The primary portion of the CHASM reporting process takes the form of a manually completed form, filled in by the responsible surgeon. Upon receipt by the CHASM team, the surgical procedure and event information is manually coded into the Read codes (which are a precursor to the SNOMED CT[4] classification) before progressing to an iterative peer review process. Adverse event causality and deficiencies of care are separately classified by 'w-codes', originally implemented by the Scottish Audit of Surgical Mortality [39]. This coding system is used at the state level in all Australian states and in New Zealand, allowing regional aggregation and reporting; however, it is not in common use elsewhere – Read codes having been superseded by SNOMED CT in most contexts and w-codes having never had broad uptake.

It would be natural to assume that an ontology aligned with adverse event reporting would be the obvious choice to meet the goals of this project, however, this is an overly simplistic view. Specifically, adverse event reporting assumes that an event has taken place, rather than targeting a risk-based approach or facilitating the detection of same. The granularity of an adverse event report itself is likely to be simultaneously too high (requiring judgements of causality and probability of recurrence, which are unavailable) and too low (losing many of the sub-procedural details, which are available for coding).

2.3.4 Provision of care, research and surveillance

Although the provision of care seems to be somewhat separate to the purposes of research and surveillance, the nature of the knowledge being handled and observed is the same. The goal of research and surveillance efforts can generally be viewed as the observation of the provision of care data. As such, these purposes will be analysed jointly in this chapter.

For clinical purposes there are a number of mature ontological models that provide significant additional meaning over the administrative models already discussed. One notable example is SNOMED CT, which represents diagnostic and therapeutic information as a combination of concepts, defined in a strict hierarchy. From this hierarchy, it is possible to derive relationships and knowledge that is not evident in the equivalent atomic unit. For example, the procedure Coronary Artery Bypass Grafting (20150531) falls under a number of taxonomic (is-a) classifications, two of which are described in Figure 6. It also has a method attribute (bypass) and direct procedure site (coronary artery structure).

[4]Systematized Nomenclature Of MEDicine - Clinical Terms.

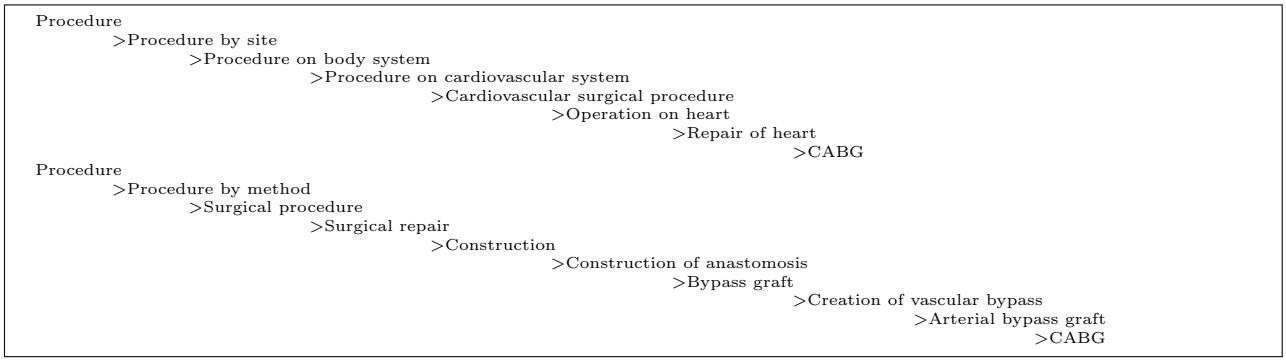


Figure 6: SNOMED CT hierarchy for coronary artery bypass graft

These classifications allow the application of logical relationships, restrictions and properties to be applied at a number of levels – e.g. a property of all blood vessel procedures, or all surgical repairs. This facilitates the grouping of elements across non-obvious lines, which would require deep domain knowledge to develop manually. It also improves information-retrieval results by the same mechanism.

Both the ubiquity of SNOMED CT and the availability of mappings into other international standards make it a good candidate for representing surgical procedures for the purpose of surveillance. It was, however, initially designed for purposes other than surgery classification (by pathologists) and therefore is expected to lack some of the sub-procedural detail that is required to fully describe the surgical reality.

As the name implies, the International Classification of Diseases (ICD) was originally designed by the WHO as a diagnostic tool. The goal of this standardisation was twofold: to make diagnostic information available for computerised analysis; and to facilitate meaningful reporting of health and mortality statistics globally.

SNOMED CT and ICD are by far the most ubiquitous clinical ontologies in use today. Table 3 provides a comparison of some important considerations for determining how these models should be implemented.

An important distinction to be made at this point is the typical primary purpose of the two models. In practice, SNOMED CT is designed for input (its high level of granularity facilitates data entry as concepts can be matched precisely and unambiguously), whereas ICD prioritises output (with an epidemiological and reporting focus, it targets meaningful aggregation and categorisation). This favours SNOMED CT for the purpose of extracting meaning from natural language, as it more closely approximates a data entry task – higher granularity terms provide more options upon which to match extracted concepts. In addition, a complete official mapping from SNOMED CT to ICD-9-CM and ICD-10 is maintained and released biannually by IHTSDO [40]. Therefore implementation of SNOMED CT provides access to the benefits of ICD as well, including compatibility with ICPS.

The GALEN[5] project on the other hand, has created an expressive formal concept model, which has been successfully demonstrated to be appropriate for the representation of surgical procedures [41]. A collaborative effort across a number of European classification

[5]Generalized Architecture for Languages, Encyclopedias and Nomenclatures.

Table 3: Comparison of ICD and SNOMED CT ontologies

	SNOMED CT	ICD
Initial Purpose	Classification of pathology (SNOMED RT) and General Practice patient characteristics (Read Codes)	Diagnostic tool for epidemiology, health management and clinical purposes
Current Primary Purpose	Data entry	Data aggregation, reporting and reimbursement
Structure	Poly-hierarchical	Mono-hierarchical
Scope	>311,000 active concepts	>155,000 active codes (ICD-10)
Maintenance Ownership	International Health Terminology Standards Development Organisation (IHTSDO)	WHO
Licensing	Royalty-free licences available throughout IHTSDO Member Territories (including Australia). Paid and/or fee-exempt licences available in non-Member territories based on intended use and income of country of use.	Non-commercial/research, internal and commercial licences available at an organisational level.

Table 4: Paradigm shifts proposed by the GALEN model

Traditional classification models	GALEN based model
Select from a fixed set of codes at the point of data entry.	Descriptive conditions from which codes can be generated when required.
Enumerated codes – similar to a phrase book where each concept has an individual definition that cannot be broken down.	Composite descriptions – more like a joint dictionary and grammar from which an indefinite number of sensible definitions can be built.
Standardisation of coding system locks code sets to their initial designed purpose, restricting flexibility.	Standard reference model allows a common means of representing coding and classification systems in order to provide interrelations.
Static coding system.	Terminology as a service – terminology servers provide a standardised software interface with which numerous applications can communicate.
Terminologies that are intrinsically linked to the language in which they are defined, leading to monolithic translation efforts.	Decouple underlying concepts from the natural language in which they are presented to provide truly multilingual systems.

centres expanded the GALEN CORE[6] model in order to provide the basis required to model surgical procedures.

One of the guiding principles for the GALEN project is that other systems are limited in that they have been developed for specific purposes (typically statistical or epidemiological in nature), and of insufficient flexibility of representation. They are therefore rarely able to be successfully applied to a goal other than that for which they were developed [42].

The solution applied by the GALEN development group in order to address these issues is to promote a series of paradigm shifts that change the way that new classifications are built, rather than simply a new classification system (detailed in Table 4).

This approach reflects the overarching trend in the design of clinical terminological systems towards reusable, modular systems, delivered as terminology servers rather than static reference lists. Older systems tend to be flat, independent and have comparatively limited reusability [43].

[6]COmmon REference.

GALEN is authored in the language GRAIL[7] and the compiled model is also made available in OWL[8]. The SPET[9] application, which is available as part of the ClaW[10] toolset, is designed specifically for creating dissections of surgical procedure descriptions (or rubrics) that conform to the expanded GASP[11] model.

It should be noted that GALEN itself is not a classification that can be implemented directly; rather, it provides an architecture and methodology upon which interoperable classifications can be built. It is in fact possible to create post-hoc dissections of the linguistic expressions found within other encoding systems in order to formalise their definition. This methodology was applied in the design of the French CCAM[12] coding system for encoding surgical procedures – a traditional domain expert consensus was achieved around the expressions available in the systems being replaced, before creating the dissections in GRAIL [44].

By providing a strictly formalised and compositional underlying model (third generation system) the GALEN models are able to derive meaning not only by groupings and classification such as in the SNOMED CT example above, but even to the point of understanding the differences (and transforming as required) between related concepts such as *viral hepatitis* versus *hepatitis virus* [45].

To take the same example of Coronary Artery Bypass Graft and model it under GALEN-CORE, it can be clearly seen that there is a much stronger typing of elements, relationships and actions, where each individual sub-unit of the classification behaves in predefined ways (see Figure 7). The detailed rules allow flexibility, as users can leverage the compositional nature of GALEN elements to describe concepts that have not been formally defined, whilst being restricted from creating nonsensical definitions, such as a fracture in a location other than the skeletal system [46].

```
SurgicalConstructingProcess
        and (Locative Attribute some
                (BypassStructure
                        and (isSpecificPhysicalMeansOf some
                                (Bypassing
                                        and (actsSpecificallyOn some CoronaryArtery)))))
        and (hasPhysicalMeans some VeinAllograft)
```

Figure 7: GRAIL definition of coronary artery bypass graft

A key benefit of such a compositional system is that the number of formally defined components or facts required to represent the domain is significantly fewer than the number of semantically compatible concepts that can be classified. It also reduces the likelihood of redundant and/or contradictory definitions, as the set of terms in the model is much smaller and therefore more manageable. This is demonstrated by the CCAM encoding, which is based on a GALEN formal representation – 7,478 procedures are represented using just 2,400 concepts and 59 semantic links [43].

[7]GALEN Representation And Integration Language
[8]Web Ontology Language
[9]Surgical Procedure Entry Tool
[10]Classification Workbench
[11]GALEN model for Surgical Procedures
[12]Classification Commune des Actes Médicaux

This compositional nature also has advantages for the purposes of translation, as the usage of each linguistic component is better defined and therefore translation of each combined term is not required. This has been shown to have benefits for natural language generation. In a study of input versus generated expressions, one third of the significant differences detected were due to errors in the definition of the dissections and two thirds were due to ambiguities in the natural language generation [44]. This is an important tool for checking consistency, veracity and integrity of a newly defined (or newly translated) model, which is not available when using traditional methods. It is possible that this will also present an advantage for automated text processing, as the target structures can be fitted to observed text more closely.

2.3.5 Automation/surgical assist systems

For the purpose of automation, surgical process modelling definitions interpret the surgical workflow as a business process. Depending on the level of granularity that is required for a given task, the components modelled may be actors, procedures, phases, actions, steps/sub-steps or motions [47].

These systems typically have their basis in interpreting signals from the OR via positional or imaging tools which observe the procedural flow of a surgical procedure – the move towards a fully integrated and 'aware' operating system is generally anticipated as more signals are made available via smart tools and systems [48].

Representations created for the purpose of automation tend to have their statistical foundation in the temporal plane [48, 49], i.e. what is going to happen next? This may be useful when monitoring aberrations from the typical workflow as an indicator or predictor of adverse events. This project is, however, looking to unlock the clinical inferences of physicians, which may include why things happened, and observations that were not available to cameras or computer enabled tools, not just detection of what occurred. As such, automation knowledge models are unlikely to be the best candidate model to apply in this instance.

2.4 Ontology selection

Based on the analysis performed in this chapter, the GALEN model is seen to be a promising target for the surgical domain. Its appropriateness has been demonstrated by its use in the French coding system CCAM. It is thus likely to meet domain-specific requirements that fall outside of the scope of other more general models.

GALEN-based models are compatible with the CEN/ISO[13] 1828:2012 standard (Health Informatics – Categorical structure for terminological systems of surgical procedures) [50] and its availability in OWL means that it also has significant compatibility with ICPS concepts.

Its linguistic formalism and compositional nature may also be beneficial in the natural language processing domain. The ability to generate natural language from the formally

[13]European Committee for Standardization (Comit Europen de Normalisation/International Organization for Standardization)

defined concepts allows for error checking when creating representation by comparison of input and generated text.

SNOMED CT AU (or at least its predecessor, the READ codes) is in use for the encoding of surgical concepts within both NSW and Australia-wide patient safety and surgical mortality audit systems [39]. It is also secondarily compatible with ICPS, via a mapping to ICD.

It is also well suited to information extraction tasks relative to other models due to its demonstrated prioritisation of data entry. In order to maximise potential interoperability, and in absence of any clear contraindicating factors, it will also be considered for its suitability as a target ontology for surgical characterisation.

2.5 Textual analysis of surgical notes

Taking the current MBS encoding system (billing) as the 'base' pre-discharge representation, this next section provides a thorough gap analysis that compares data available within the operative report to the coded information that is accessible for automated surveillance. The goal of this activity is to help prioritise information extraction activities – those target data points that are less well covered may be more valuable in extraction (assuming clinical implications are equivalent).

This will then be compared to the coverage provided by a manual encoding into the two selected target representations of SNOMED CT and GALEN, in order to further test appropriateness for the representation of surgical procedures.

2.5.1 Gap analysis: methods

A corpus of operative reports was obtained from MUH that represents a mix of surgical specialties.

The de-identified data set obtained from MUH contained data from 861 patient records. The data collected comprises:

- Surgical notes and observations.
- MBS codes assigned to the procedure.
- Some minimal administrative data points.

In order to systematically identify all gaps in current structured data representation, 50 surgical notes were randomly selected from the available set. Each of these notes was first annotated and then compared to the concepts available in the MBS data, and then manually encoded into both the SNOMED CT AU and GALEN models, with the intention of quantifying the adequacy of the target structures identified above, recognition of any potential gaps, and comparison with the status quo.

2.5.2 Annotation guideline

For the purpose of defining the base representation required for gap analysis, annotations of operative reports were prepared by identifying the following categories of report elements. This list was derived by reviewing the five longest (and thus assumed most detailed) reports by word count and defining categories which encompass all concepts within.

- **Action:** CEN ENV 1828 defines both surgical deeds and procedures – the former being a surgical action that can be performed to a patient's body, irrespective of location (e.g. *cut, debride*) and the latter being specific to a location (e.g. *colonoscopy*). In this context, both of these subgroups are combined into the concept of a surgical action.
- **Anaesthesia:** Details of anaesthetic techniques or agents listed within the operative reports e.g. *GA, local anaesthetic, nerve block*.
- **Anatomical location:** This is the body system or structure where the surgical action is applied. This includes any anatomical location in the adjectival form such as *umbilical, epigastric*.
- **Approach:** The method of approach for the surgical action. This may be either a technique such as *robotic*, or *laparoscopic* versus *open*, or mode of access such as *intracranial* or *transsphenoidal*.
- **Closure:** Any information provided on the method of closure for wounds, including *method, location, length, stitch count* and *types*.
- **Defect target:** In the case of a surgery involving a repair, removal or other operation specific to a pathology, this is the identified target. This could be a *tumour, tear, hernia, occlusion* etc.
- **Device**: A surgical device (as opposed to tool) is a non-reusable item that remains in the patient after surgery. This includes items such as *screws, stents, grommets, pacemakers* or *replacement joints*.
- **Finding:** A diagnostic observation made during surgery. This may be evidence of disease (or absence of), or the size and type of observed pathology, e.g. *5cm malignant tumour*.
- **Finding location:** The anatomical location of a finding, as differentiated from the anatomical location, or target of the surgery.
- **Locative:** A locative concept with respect to the target anatomy may be unequivocal (*left/right, anterior/posterior, superior/inferior, medial/lateral, proximal/distal, bilateral*) or mentioned but ambiguous (typically in the coded information as 'either side'). It is also possible to derive an unambiguous location from multiple instances of an otherwise ambiguous code – e.g. two instances of the code *41752*, a code for an *intranasal operation on the sphenoid sinus* that does not have a specific laterality, implies a bilateral surgical procedure.
- **Positioning:** Description of positioning of the patient or body-part during the procedure, e.g. *lithotomy, rotated*.

- **Purpose:** An insight to the clinical rationale behind a surgical procedure, e.g. *exploratory, diagnostic.*
- **Tool:** This is any tool (reusable or consumable) that is required to perform the surgery but is not left in the patient at the conclusion of the procedure.

For each report, all mentioned concepts were categorised. This was repeated for the textual definition of each associated MBS code. A concept was defined as the smallest logical unit that preserved the intended meaning of the phrase. For example, *pars plana* indicates an anatomical site; however, in the context *via pars plana sclerotomies*, it is labelled jointly with the word *via* as an approach for the action *sclerotomy.*

Cross-referencing was then performed for each concept, which was consequently classified as *covered, ambiguous* or *missing.* If a concept was repeated identically (word for word, and in the same context) in two separate sections of the report, the repeated text was excluded from analysis.

If a concept was available in the coded data, but not the written report, this was taken as assumed background knowledge for a given procedure and not counted within the classification.

For each concept that could be unambiguously matched from report to the coded data, this is *covered.* For a concept that is contained in both report and coded data, but could not be explicitly matched, this is *ambiguous* – typically this occurs when a code covers more than one type of procedure e.g. *49562: Knee: arthroscopic surgery of, involving 1 or more of: partial or total meniscectomy, removal of loose body or lateral release...*, or where the written report is more specific to subtype than the coded data, e.g. *Rathke's cleft cyst* versus *pituitary tumour.* A *missing* concept is one that is present in the written report but it would not be possible to derive from reviewing the coded data alone.

In the case of any unclear classifications, a guideline for coverage or lack thereof was to ask the question: could one apply the research question *'was concept <x>the target of/undertaken during the procedure'* and determine the correct answer by looking at the coded data alone? This allows matches for concepts such as *arthroscopy* (action) or *haemorrhoid* (defect target) in the report to be matched to *arthroscopic reconstruction* (approach, action) or *haemorrhoidectomy* (action) in the coded data. This was done in order to provide an accurate picture of concept coverage, as opposed to a strict semantic matching of each concept.

2.5.3 MBS annotation: results

One report was found to have been assigned incorrect billing codes – an elbow arthroscopy was linked to a single code that indicated a similar procedure on the shoulder – and was therefore excluded from analysis.

Four of the randomly selected reports contained concepts that could not be classified into the aforementioned groups – two mentioning that the operation was a repeat or revision of a prior operation, one which included a family history note, and the other listing pre-operative pain characterisation. This very low miss-rate (1%) implies that the list of categories

chosen for this analysis is indeed representative of the information that is typically included in operative reports at MUH.

The coverage of concepts within cardio-thoracic reports is somewhat of an outlier (see Table 5). This is due to the inclusion of a single procedure, defined by very general angiography procedure codes – neither specific to location nor technique. This is not the only report of this nature – in fact, three reports contain no precisely coded elements, however its effect is highlighted by the low number of cardio-thoracic procedures selected.

Table 6 shows the coverage properties overall, and by concept type. An operative report that is representative of the annotated sample contains an average of 8 concepts (mode=6, s.d.=3.8), 2.4 of which (mode=2, s.d.=1.9) can be unambiguously determined from the billing codes. This high level of missing or ambiguous data in the MBS codes (48% and 21% respectively) demonstrates clearly that the use of billing codes for the purposes of surveillance or research is limited and that unlocking these free text reports for analysis is a worthwhile goal.

The concepts that consistently demonstrate relatively high levels of coverage (action, anatomical location and approach) are intuitive when considering that the purpose of billing codes is to capture reimbursable actions.

Defect target concepts have patchy coverage, implying that the type of defect that is targeted by a given procedure is only sometimes relevant to the reimbursable nature of the procedure – e.g. codes often describe a procedure on a *tumour*, where reports include more exactly which type of tumour such as *macroadenoma*, or codes that mention a *repair*, without describing the associated defect at all. This detail may be clinically relevant, and thus important for calculations of risk, predictive models or determination, however has no bearing on payments made to the clinician.

Devices have extremely poor coverage in the coded data; however, in this particular hospital, procedures dictate that all devices are recorded elsewhere in the medical record via barcode scanning and therefore their derivation from billing codes is not required. Findings, and finding locations are also likely to be included in diagnostic codes elsewhere in the medical record.

Locative concepts also have very low levels of coverage in the MBS codes. Again, this is somewhat intuitive – it is rare that it would matter for the purposes of payment whether the left or right side of the body was the target of the operation. Indeed, most of the instances of 'covered' locative items were for bilateral procedures, where it was clear from the inclusion of two instances of the billing code that this is what had occurred. In the majority of procedures, this is also unlikely to make a clinical difference; however, there are certain instances where it will be relevant. Consider the following distinctions:

- Division of adhesions in the colon in either the ascending (right) colon or descending (left) are represented by the same MBS code *(30393)*.
- Repair of a distal or proximal dislocation of the radio-ulnar joint *(47027)*.
- Repair of the medial or lateral collateral knee ligament *(49503)*.

Closures have no coverage whatsoever within the billing code data. From the types

Table 5: Annotated report sample by specialty

Specialty	Count	Total concepts	Covered (%)	Av. concepts / report
Breast Surgery	1	8	50.0	8
Cardio-thoracic Surgery	1	7	0	7
Colorectal Surgery	6	38	44.7	6.3
Gastroenterology	4	39	30.7	9.75
Gynaecology	1	5	20.0	5
Hand Surgery	4	36	27.8	9
Head and Neck Surgery	6	36	30.6	6
Neurosurgery	8	66	21.2	8.25
Ophthalmology	6	45	33.3	7.5
Orthopaedic Surgery	10	102	33.3	10.2
Urology	2	11	18.2	5.5
Total	**49**	**393**	**30.5**	**8.0**

of closure information observed in the selected reports, it is unlikely to have primary clinical significance; however, it may be possible to derive a secondary measure that can be associated with risks or outcomes such as number of stitches or type of stitches used as an indicator of size and complexity of the surgical wound.

All other concept types (anaesthesia, positioning, purpose and tool) occur with too low frequency within the sample to make any comment on their coverage or otherwise.

2.5.4 SNOMED CT and GALEN encoding: results

Assuming the MBS encoding of approximately 31% unambiguously available concepts represents a minimum from which any proposed knowledge model must improve, manual encoding of the experimental set of 49 records into both SNOMED CT and GALEN was performed. The same annotations were used as detailed in the previous section.

The encoding of operative reports as SNOMED CT or GALEN models has significantly higher coverage than observed when applying billing codes alone. This again is intuitive, given the different purposes for which these ontologies were produced. This manual

Table 6: Concept category frequencies

Category	Count	Covered	Ambiguous	Missing
Action	114	54 (47%)	37 (32%)	23 (20%)
Anaesthesia	2	1 (50%)	0 (0%)	1 (50%)
Anatomical location	64	29 (45%)	17 (27%)	18 (28%)
Approach	17	9 (53%)	3 (18%)	5 (29%)
Closure	21	0 (0%)	0 (0%)	21 (100%)
Defect target	44	16 (36%)	12 (27%)	16 (36%)
Device	29	1 (3%)	9 (31%)	19 (66%)
Finding	32	2 (6%)	1 (3%)	29 (91%)
Finding location	13	2 (15%)	1 (8%)	10 (77%)
Locative	55	6 (11%)	3 (5%)	46 (84%)
Positioning	1	0 (0%)	0 (0%)	1 (100%)
Purpose	0	0	0	0
Tool	1	0 (0%)	0 (0%)	1 (100%)
Total	**393**	**120 (31%)**	**83 (21%)**	**190 (48%)**

(Note: percentages may not add to 100 due to rounding.)

activity is intended to provide a basis for understanding on which to build an automated classification system, rather than to identify specific gaps or limitations as per the previous MBS code analysis.

Within the set of included operative reports, only 16 reports contained one or more SNOMED CT uncodable concepts (22 total uncodable concepts) and 20 reports had at least one GALEN uncodable concept (28 total). This 5-7% miss rate clearly demonstrates the value of targeting a knowledge model that is more fit for purpose than billing codes. This will provide a far more reliable basis for the design and implementation of automated surveillance tools when compared to the currently available encoding.

2.5.5 SNOMED CT and GALEN: full evaluation

As both SNOMED CT and GALEN have been demonstrated to closely align with the target information from free-text operative reports, and since there is a very small difference in the level of coverage provided for the available information, they will both be subjected to the full evaluation outlined previously.

Table 7: Full evaluation of SNOMED CT and GALEN ontological models

	SNOMED CT	GALEN
Completeness, Richness and Granularity	Outperforms ICD-9-CM and ICD-10 (among others) in every tested category across the clinical record, with the highest level of difference (above both the average of all measures and average of ICD measures) seen in the modifier (negation, size, severity etc.) and treatment and procedure categories [51].	No formal study of the coverage provided by the GALEN architecture was found; however, as demonstrated in the previous section, its coverage of concepts observed in operative reports approaches that provided by SNOMED CT. There was also no clear distinction between categories of concepts that are not covered by each ontology.
Conciseness	An algorithmic review of SNOMED CT (2003) [52] found a conservative estimate of 3% redundant concepts. It can be presumed that later versions include the retirement of some of these redundant concepts; however, it is also likely that new redundancies have been introduced.	GALEN prioritises the minimal ontological commitment more highly than SNOMED CT, which prioritises instead the convenience of description and computation [53]. This is more 'correct' in the strict ontology modelling sense, however in practice may be found to be of limited or even detrimental value.

Consistency, Accuracy and Coherence	SNOMED CT has been found to contain numerous errors in definition of pre-coordinated concepts, e.g. [52, 54]. This must be expected for a manually defined terminology of this scope. Post-coordinated concepts rely on these existing definitions and thus can be assumed to cascade these same limitations. Users can, however, have relative confidence in the consistency, accuracy and coherence of the pre-coordinated concepts in SNOMED CT – these concepts have undergone many levels of review, and having been widely implemented for more than a decade it can be assumed that errors with clinical significance have been reduced dramatically.	GALEN contains similar errors of definition [55]; however, due to far lower adoption levels, there is much more limited analysis of these than for SNOMED CT. Likewise, with a less active community and maintenance structure, the detection and correction of these are less likely. GALEN also relies more heavily on post-coordinated definitions, and therefore errors may take longer to surface.
Expandability and Adaptability	Given the high level of uptake globally, the level of resource that is allocated to maintaining and authoring SNOMED CT concepts is high relative to GALEN. As SNOMED CT is a formally and actively managed terminology, it has been undergoing continual updates since its initial release in 2002. The SNOMED International Request Submission System provides facility for end users to propose update/addition/retirement of concepts.	The lack of wide adoption of the GALEN architecture, and no evidence of a publicly available path to propose modifications implies a very sluggish expansion and error resolution process.

Clarity	SNOMED CT has a relatively well-established synonym list for each concept, which improves usability for human users. Additionally, its prioritisation of data entry tasks favours common-usage versions of terms, which increases intelligibility.	GALEN has been demonstrated to facilitate meaningful natural language generation [41]. Despite the modelled concepts being difficult to parse directly, the generation of natural language text in multiple languages expands the audience who may utilise and modify these concepts easily. In the absence of natural language generation capability, the intermediate notation has also been shown to provide clarity and ease of understanding [56].
Computational efficiency	It has been demonstrated that even with a simplistic algorithm that does not take into account all of the richness of the SNOMED CT classification, a meaningful semantic similarity can be calculated between terms, based on path length, information content and context vector [57]. Due to the significantly higher number of terms, SNOMED CT is likely to suffer from some computational inefficiencies when compared to GALEN.	It is theoretically possible to compute semantic similarity using GALEN; however, no implementation of this was found, so it is not possible to provide comment on the efficiency or otherwise. The exclusively post-coordinated nature of GALEN lends efficiency to the implementation of logical inference and query.
Organisational fitness	SNOMED CT is freely accessible to Australian hospitals, as an IHTSDO member state. Its available cross mapping with ICD-10 and thus compatibility with ICPS also makes it a good organisational fit for the purpose of safety monitoring and surveillance.	GALEN is an open source model and thereby available to all potential stakeholders. A representation of ICD-10 under the GALEN architecture has been previously attempted and rejected due to difficulties of representing all anatomical concepts [58] – this lack of compatibility is a cause for concern with respect to patient-safety specific goals.

In a purely ontological sense, the GALEN model provides a richer and more flexible representation of surgical procedures than SNOMED CT. Based on the full evaluation,

however, it is clear that the broader uptake of SNOMED CT has significant advantages in terms of compatibility, expandability and accuracy. As the performance in the test encoding is otherwise very similar, SNOMED CT will be selected as the target ontological model for this project. It has been demonstrated to have sufficient coverage of concepts available in operative reports, and compatibility with adverse event reporting mechanisms.

The next chapter will provide a review of technologies and methods that are required to move toward an automated encoding of operative reports, and the final chapter will apply these methods to work towards the extraction of SNOMED CT concepts.

3 Natural language processing in the surgical domain

3.1 Background

The selection of an idealised model to adequately represent surgical procedures may be a valuable enterprise, however in practice it is unlikely to be successfully integrated with existing systems and workflows unless its encoding can be automated.

Health information systems (HIS) have been introduced into the clinical setting at a rapid pace, rising from less than 10% electronic health record (EHR) adoption in 2008 to over 40% in 2012 (US figures) [59], and electronic billing systems have been sufficiently prevalent so as to be considered ubiquitous as early as 1990 [60]. These systems contain vast quantities of data detailing patient histories, diagnoses, clinical notes, interventions applied, test results, vital signs and billing, which present significant opportunities for the monitoring of adverse events and assessing patient outcomes.

Despite these clear opportunities, the use of these systems for the purpose of monitoring or improving patient safety remains ad hoc due to the numerous challenges presented for their systematic implementation [61] – most notably the use of free text fields for many clinically significant notes, sparse and inconsistently coded data sets, high costs of manual reviews where required and the use of diverse systems which lack interoperability, leading to incomplete data and challenging longitudinal follow-up.

Patient safety systems represent a subset of health information systems that are used for the purpose of surveillance, prediction or improvement of patient safety outcomes. Each of these activities is inherently rooted in data analysis and therefore their effective design and generalisability across systems is closely linked to the collection methods, quality and accessibility of the data underlying their design.

It is therefore necessary to create patient safety systems with the ability to accurately and intelligently interpret this unstructured text in an automated fashion. The use of natural language processing (NLP) techniques to this end has been applied with varying levels of success. The remainder of this chapter will first give a brief overview of NLP techniques and the way in which they are typically applied within clinical systems, before delving more deeply into the statistical properties of the text within operative reports, and finally present a prototype algorithm for extraction of key concepts.

3.2 General natural language processing techniques

NLP involves the use of automated procedures to detect concepts and extract non-trivial knowledge from unstructured text. A typical NLP system will often contain a number of distinct sub-tasks, which are chained together to form a processing pipeline of sorts. This is done in order to be able to handle the complexities of language that make straightforward dictionary matching incapable of handling any but the simplest data extraction tasks.

Table 8: Clinical NLP systems

System	Initial Release	Primary Target	Technology	Corpus
LSP: Medical Language Processor [62, 63]	Project: 1965, clinical specialisation: 1987	Discharge summaries, progress notes, radiology reports	Sublanguage theory, Structured Health Markup Language (SHML)	No statistical methods – lexicon derived from publicly available sources and clinical records
MedLEE [64, 65]	1994	Radiology reports, later expanded to mammography, discharge summaries, electrocardiography, echocardiography, pathology	Semantic grammar, with limited syntactic rules	No statistical methods – lexicon based on UMLS, with additional terms drawn from clinical terminologies
SymText [66, 67]	1994	Chest x-ray reports	Augmented transition network grammars (syntactic) and Bayesian networks (semantic)	Training documents (583, 3152 and 3152 documents for: appliances, diseases and findings respectively) of unclear origin [68]
cTAKES [69]	2006	General clinical notes	Modular system combining rule-based and machine learning components	Mayo Clinic EHR corpus – 273 reports, of mixed type, in addition to the publicly available PTB and GENIA corpora (non-clinical/general text and biomedical respectively)

HITEx [70]	2006	Airway disease (asthma, smoking habits) in discharge summaries	Modular system combining rule-based and machine learning components	150 discharge summaries containing asthma or COPD-related billing codes, obtained from Bringham and Women's Hospital
MedKAT/P [71]	2009	Pathology reports, cancer-specific characteristics	Modular system combining rule-based and machine learning components – based on OpenNLP components with domain-specific modifications, Cancer Disease Knowledge Model	302 training documents (201 training, 101 test) of unclear origin

This processing pipeline makes use of a number of machine-readable knowledge sources, including dictionaries, thesauri, rules of grammar, linguistic properties of words, statistical models and an ontological model of the target information.

NLP applications vary, with differing levels of complexity, knowledge sources, techniques and permutations of sub-tasks, as required by the source text and target extraction. Figure 8 describes an example implementation containing many of the common NLP pipeline components. Systems have been proposed to absorb these subcomponents into deep neural networks that may be jointly trained, eliminating this pipeline structure [72], however it remains typical at the point of writing.

In early implementations of NLP, the development focused on the definition of formal grammars by linguists, with the goal of finding a representation providing complete coverage of the source domain. This is, however, impractical for general text as the number of rules and exceptions becomes vast, and true ambiguity persists despite the increase in granularity. It also tends to be incapable of digesting the highly abbreviated and ungrammatical text prevalent in clinical systems [73].

Most modern implementations have moved away from this strictly formal analysis and tend to apply instead a statistical approach. This has the advantage of being able to explain the levels of uncertainty and incompleteness that commonly characterise the phenomena present in natural language [74]. This means that statistical methods are robust, generalisable and behave more gracefully than rule based or heuristic methods when presented with data that has not been seen previously.

3.3 Clinical domain specific resources

Many groups have developed systems tuned for biomedical texts, but a smaller amount of work has been done on systems specifically tuned for the clinical domain. Table 8 provides an overview of some key tools that have been released in this space.

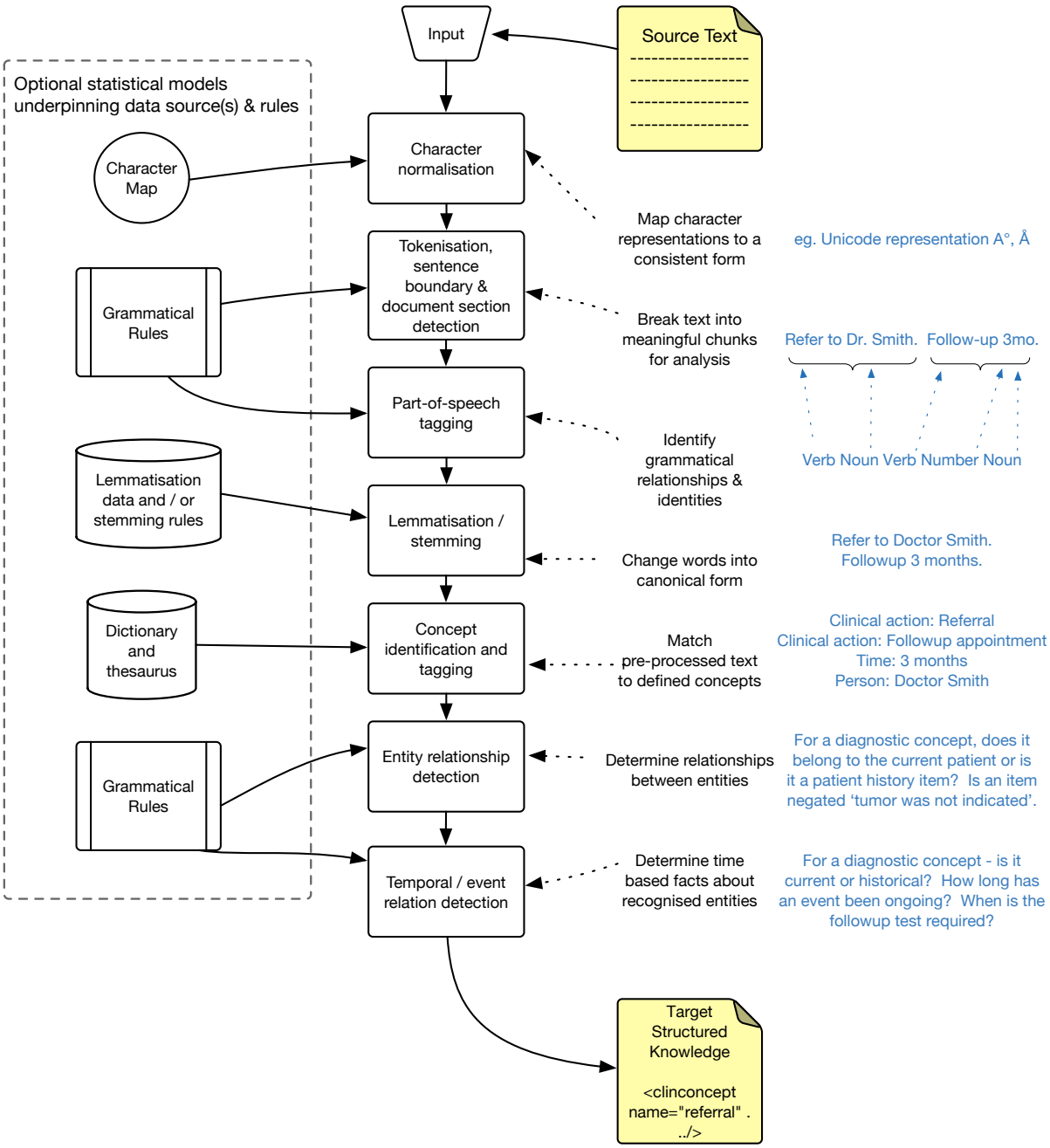
3.4 Domain challenges

A statistical approach to NLP requires significant effort to create large, annotated corpora in order to provide a basis for observation.

Anecdotally, clinical text is characterised by a high number of abbreviations, wide-ranging vocabulary and idiosyncratic, staccato grammar. The use of generally available corpora such as the Penn Treebank [75] is therefore unlikely to be successful when applied to clinical NLP tasks.

It is challenging to curate a corpus of this nature for clinical text, as patient confidentiality issues restrict the availability of meaningfully large sets of training data. The capacity of groups to share data is limited for the same reason. Referring to Table 8, the only data available for public use from this set is released by i2b2, who are behind the HITEx application [76].

Figure 8: Example NLP pipeline



Operative Diagnosis: Cholecystolithiasis

Operations Performed:

MBS Item No(s)	Operation Name
30445	Cholecystectomy, laparoscopic
30439	Operative cholangiography

Details of Operation:

thin walled GB encased in fat
multiple small pigment GS
cholangiogram satisfactory

Subumbilical Hasson port
2X R sided 5mm ports
L-epigastric 100m port
cystic duct and artery skeletonized
transcystic cholangiogram
GB resected. Inadvertent perforation / spillage of GS
generous saline wash
spilt GS removed with irriagtion and suction

Post Operative Orders:

RPAOs,
Sips O/N as tolerated,
Analgesia,
S/C heparin,
TEDS

Figure 9: Sample operative report

In addition, the corpora underpinning the tools in this table are relatively small when compared to biomedical or general domain corpora, e.g. the Colorado Rich Annotation of Full Text corpus (CRAFT) at 790,000 tokens (biomedical) [77] or the Penn Treebank with 4.5 million words (general) [75].

It is not possible to reproduce an individual operative report from the experimental data for illustrative purposes due to patient confidentiality requirements. A freely available example from a RACS-approved audit program has been provided instead [78] (see Figure 9 – N.B. spelling and grammatical errors preserved, as these will be relevant to any processing algorithms applied).

This sample was selected because it is of a similar nature to many of those observed in the MUH data set. By simple inspection, it is clear that the type of language that is used in an operative report is very different to that which is used in ordinary communication, or even in biomedical articles. There is likely to be variability of practice between institutions, regions, clinical specialties and even at the per-clinician level; however, it is expected that the cohesion of the operative language will be higher than its similarity with other texts.

A study of the use of verbs within operative reports [79] demonstrates that the sublanguage employed is clearly distinct from not only standard English-language text, but also other types of clinical text – only 11.5% of verbs in the operative reports studied were found to be defined within the domain-specific lexical resource UMLS.

Under the assumption that neither the vocabulary nor grammar of these reports are typical of written language, it is expected that the text will also follow the sublanguage behaviour as observed in subdomains of scientific literature [80]. This means that it will be possible to observe patterns of subsets of vocabulary appearing in specified grammatical relation to one another. Defined statement types belonging to a sublanguage may be nonsensical in general language (such as the high number of sentences that do not contain a verb seen in Figure 9). These would be vanishingly unlikely in biomedical texts and newspaper articles, and highly unlikely in more informal texts such as emails; however, are common in the context of an operative report, and have the clear contextual implication of a temporally sorted

list of activities describing the surgery.

When applying statistical NLP methods, the impact of the presence of a sublanguage for clinical text, and more specifically, operative reports, is threefold [81]:

- Syntactic regularity: Classes of words can be found to behave in expected ways relative to one another within documents of a defined sublanguage. This is relevant for relationship and anaphora resolution, as it becomes possible to predict the role of each word within a given phrase. This may help in word sense disambiguation, where a term or abbreviation has more than one definition, and the correct interpretation is provided statistically by context. In the instance of spelling mistakes and atypical abbreviations, it can also allow accurate replacement of terms by the appropriate canonical form. In the example report provided (Figure 9), the verb *skeletonised*, which is uncommon in standard English text but describes a commonly employed surgical action, could be identified as an action due to its position and relationship to other terms in the sentence*Cystic duct and artery skeletonised*.

- Inequalities of likelihood: Within a specific sublanguage, terms have a certain probability of occurring as arguments of a given operator. See, for example, the regularity of *Subumbilical Hasson port. 2X R sided 5mm ports. L epigastric 100m port* from Figure 9. The word *port(s)* (or nouns with statistically similar usage) at the end of a verb-free sentence within the operation description section of the report increases the likelihood that the preceding terms are type or size descriptors.

- Paraphrastic reductions: If a term is so likely in a given context as to be redundant, it is considered to have zero information content. This can be clearly seen in the sentence *Inadvertent perforation with spillage of GS*. In this instance, the clinician has not felt that it is necessary to make note of which organ was inadvertently perforated – it is clear both from the surgical target associated with a cholecystectomy procedure and the contents of the spillage (gallstones) that it is in fact the gallbladder that was breached. It would be unlikely that an anaphoric relationship such as this could be resolved under general language; however, with sufficient appropriate training examples in the relevant sublanguage it may be possible to refer a cutting action without an object back to the target of the procedure itself.

One benefit of statistical approaches to NLP is that they can perform equally well in the face of grammatical or ungrammatical text, on the assumption that the target text is ungrammatical in a uniform way, i.e. the presence of a defined sublanguage. That is, limited prior knowledge is taken into account other than the text that forms the corpus – if a corpus is adequately representative of the text in question, and labelled items are accurate, the NLP application will continue to perform acceptably.

3.5 Understanding the sublanguage of operative reports

It is therefore a key starting activity to understand the nature of the text that will be processed as part of this project. In order to do this, the methodology used by Verspoor et al. [82] to compare the nature of text between Open Access and traditional scientific journals has been replicated here.

Table 10: MUHON corpus properties

Paragraph type	Count	Non-empty count
Operation Performed	901	872
Details of Operation	852	214
Operation Findings	562	472
Closure	460	453
Tubes/Drains/Caths left insitu	263	263
Post Operative Instructions Surgeon	889	665

The goal of the methods being reproduced was to validate the use of Open Access articles to produce generalisable results when developing biomedical text mining algorithms. Open Access articles are commonly used due to their accessibility and free availability; however, if there were found to be significant differences in content or format, the applicability of this work to articles published in traditional journals is limited. The authors found that the semantic and syntactic similarity of Open Access and traditional journal text was sufficiently high to confidently apply research results found in one type of article to the other.

The null hypothesis presented here is that the text in operative reports closely approximates text found elsewhere in the clinical record and thus it can be expected that clinically tuned NLP systems (such as cTAKES) can be applied directly to operative notes. The following experiment will aim to find differences that refute this hypothesis.

3.5.1 Operative report sublanguage: methods

Four corpora were assembled for the purposes of this comparison.

- **MUHON** is a set of 901 operative notes that were collected from Macquarie University Hospital between 2010 and 2015. These randomly selected operations were performed across 874 admissions and represent data from 861 patients. The text was entered by a member of the surgical team (typically the surgeon or assisting surgeon) as free text under the following synoptic headings – *Operation Performed, Details of Operation*, *Operation Findings*, *Closure*, *Particulars of Tubes/Drains/Catheters left insitu.*, and *Post Operative Instructions Surgeon.* This is comprised of 4349 paragraphs, 3245 of which are not empty, totalling 6789 words.

- **i2b2 Discharge Corpus** [83] is a collection of fully de-identified discharge summaries, which is made freely available for research purposes by the Informatics for Integrating Biology and the Bedside (i2b2) group. The original purpose of this corpus was for an automated de-identification NLP challenge. It is included here as a comparison text collection because discharge summaries represent text drawn from a diverse cross-section of the clinical record. This will allow comparisons of the specialised operative notes to general clinical text. This will provide a basis to understand to what level existing parsers tuned for clinical text can be expected to be effective. The un-annotated data set was used for this analysis.

The discharge summaries were collected under the synoptic headings – *Admission Diagnosis*, *Principal Diagnosis*, *Associated Diagnosis*, *Discharge Diagnosis*, *Other Diagnoses*, *History of Present Illness*, *Past Medical History*, *Reason for Admission*, *Medications on Admission*, *Allergies*, *Family History*, *Social History*, *Physical Examination*, *Adverse Drug Reactions*, *Clinical Infections*, *Special Procedures and Operations*, *Laboratory Data*, *Principal Procedure*, *Hospital Course and Treatment*, *Discharge Medications*, *Condition on Discharge*, *Discharge Disposition*, *Doctor Discharge Orders*, *Follow up*, *Addendum to Discharge Summary*, *Additional Comments* or just *Preliminary Report* if the record in question was an Emergency Department admission. Minimal other data items such as *Admission/Discharge Date*, *Dictation Date*, *Report Signature Status*, *Attending/Dictating Doctor* are available in the data set but were excluded from analysis (free-text fields only were retained).

This set contains 919 reports, 97 of which contain only excluded fields. The 822 included reports contain a total of 350,081 words.

- **Reference** is based on a 5% subset of the Penn Treebank [75], which contains data from the *Wall Street Journal* and is available freely under fair-use for non-commercial purposes. This text represents general non-clinical text, and comprises 199 articles and 100,676 words.

- **PIL Corpus** [84] is a collection of text from patient information leaflets, which was created based on the Association of the British Pharmaceutical Industry compendium. This corpus aims to be representative of clinical topics such as is contained in the medical record; however, written in a style which is intended for comprehension by the general population, as opposed to the discharge summaries and operative reports created for later clinical reference. There are 474 files in this corpus, totalling 587,484 words.

Table 11 reports the number and incidence rate of a number of the morphosyntactic/semantic phenomena in the four corpora.

For the purposes of this analysis, simple heuristic algorithms were defined as per Verspoor et al. [82] (where available), or as described in the following list (including rationale for any deviations). All algorithms were implemented in Python 3.4.3, based on the Natural Language Toolkit (NLTK) [85]:

- Sentences were segmented using the `nltk.tokenize.punkt` module. This module uses an unsupervised algorithm to build a model that can take into account likely abbreviations, collocations and words that typically indicate the start of sentences [86]. For each corpus, sentence segmentation was performed using the included pre-trained English language tokenizer. Accuracy and f-score of the pre-trained general English language tokenizer model was then estimated by reviewing a randomly selected 5,000 character sample.

- Tokens were counted using the `nltk.tokenize.word_tokenize` function, which is based on the `TreebankWordTokenizer` – see for example:
```
s="Percutaneous insertion: 3x 1.4mm k-wires"
nltk.word_tokenize(s)
['Percutaneous', 'insertion', ':', '3x', '1.4mm', 'k-wires']
```

- Type counts provided are case insensitive.
- The stopword list used was `nltk.corpus.stopwords.words('english')`.
- Negation was counted as instances of the words *no*, *not*, *neither*, *nor* and the affix *n't*, per Verspoor et al., with the addition of *never*, *none*, *nothing* and *nowhere* in order to cover expected negation of patient history and observation items, such as *the patient has never...*, *nothing was observed...*, *none of the measurements...*, *nowhere on the scan...*
- The estimation of coordination, pronouns and passives were all determined as described in Verspoor et al.

3.5.2 Operative report sublanguage: results and discussion

Table 11: Incidence of syntactic/semantic phenomena – comparison between corpora

	MUHON	i2b2	Reference	PIL
Document count	901	822	199	474
Sentence count	183	20,819	3,970	30,389
Sentence count recall	99.32%	99.86%	99.94%	99.82%
Sentence count f-score	19.05%	92.47%	95.52%	89.16%
Avg. sentence count	0.20	25.33	19.95	64.11
Token count	6,570	327,520	100,918	604,645
Type count	1,759	16,239	12,048	14,057
Stopword count	837	93,190	30,474	218,530
Stopword %	11.88%	26.55%	30.27%	37.20%
Avg. document length	7.29	398.44	505.91	1,275.62
Avg. sentence length	35.90	20.17	25.42	19.90
Types/Tokens	25.0%	4.6%	12.3%	2.3%
Tokens/Types	44.0	21.6	8.1	41.8
Negatives	12	2,835	627	5,516
Negatives %	0.17%	0.81%	0.62%	0.94%
Coordination	168	8,624	1,859	16,525
Coordination %	2.39%	2.46%	1.85%	2.81%
Pronouns	42	14,586	5,142	58,565

Pronouns %	0.60%	4.16%	5.10%	9.97%
Passives	1	546	12	1108
Passives %	0.014%	0.16%	0.012%	0.19%

K-L divergence: The Kullback-Leibler (K-L) divergence is a measure of the relative entropy of two probability distributions. This can be interpreted as the expected additional number of bits required to encode a corpus c_1, using a code that is optimised for corpus c_2, rather than its own optimally devised code. This is defined for words w in the vocabulary V created by combining unique terms in c_1 and c_2 as follows (and as defined in Verspoor et al):

$$D_{KL}(c_1||c_2) = \sum_{w \in V} \left(p(w|c_1).ln\frac{p(w|c_1)}{p(w|c_2)} \right)$$

To allow for comparison, it is converted to a symmetric measure of divergence by taking the minimum:

$$Divergence(c_1, c_2) = min\left\{D_{KL}(c_1||c_2), D_{KL}(c_2||c_1)\right\}$$

This value is undefined for terms that occur in one corpus and not the other, therefore Laplace smoothing was implemented, which assumes a minimum frequency of 1 in each corpus $\forall w \in V$.

As per Verspoor et al., this was calculated for the n most frequent words in the combined vocabulary V, for different values of n, presented in Table 12. For the corpora observed here; however, the limiting factor was the type count of the MUHON corpus, therefore n was set to range from 100 to 1,500, instead of 100 to 10,000.

The disparity of size between these corpora limits the validity of calculating the K-L divergence directly – the top n words by frequency in any combined vocabulary will skew heavily to the larger corpus. As such, for the K-L measure only, a contiguous subsection was randomly selected from the larger corpora to create sub-corpora matched for word count, with the MUHON word count as the base size.

Log Likelihood: It is possible to find the terms that contribute the highest relative frequency difference, and thus can be interpreted as identifying terms for a given body of text using the log likelihood measure. It is calculated using the following equations, where E_i is the expected value for a term t in c_i, O_i is the number of occurrences of t and N_i is the number of types in c_i [87].

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

$$-2ln\lambda = 2\sum_i O_i ln\frac{O_i}{E_i}$$

The log likelihood is also undefined for instances where O_i is 0; however, since $\lim_{O_i \to 0} O_i ln\frac{O_i}{E_i} = 0$, terms where $O_i = 0$ are ignored. The top 10 most distinctive words

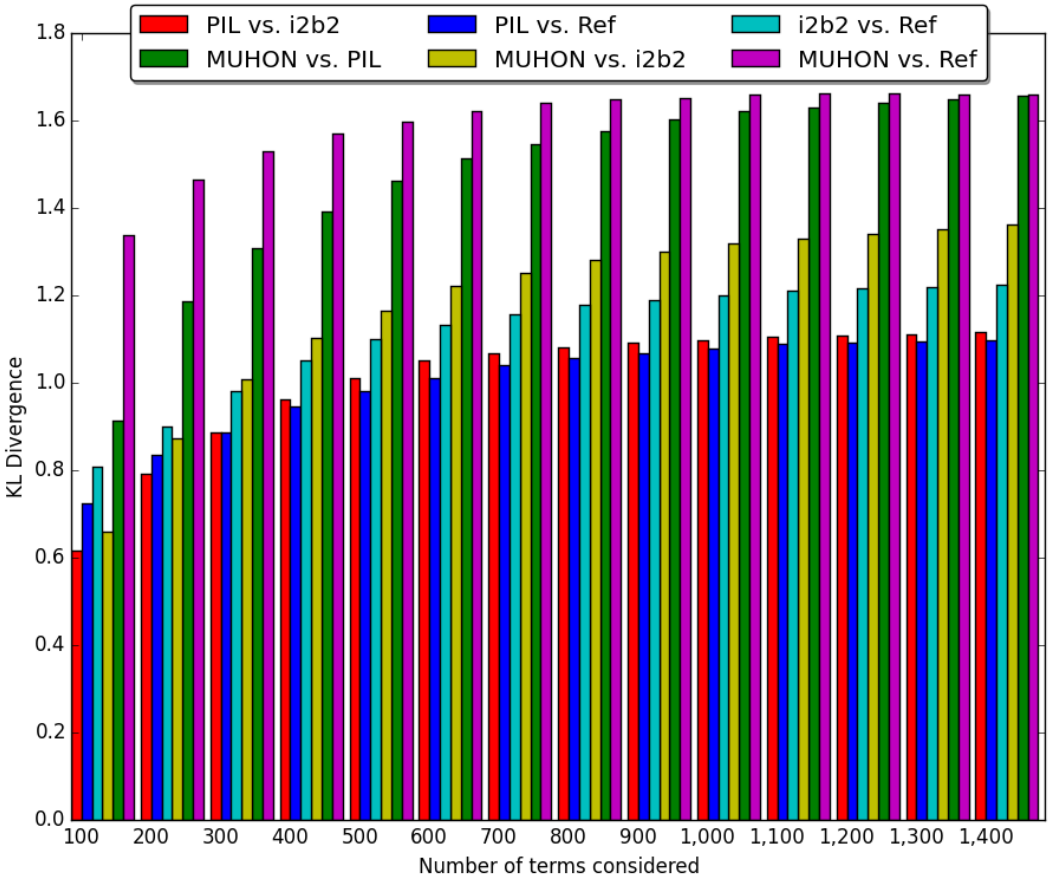


Figure 10: K-L divergence by number of terms considered

when compared to the MUHON corpus are presented for each of the other experimental corpora in Table 13.

The measures describing the MUHON corpus have the least similarity when compared with the other corpora. The most notable differences can be generalised under the following categories:

Reliability of sentence count: Sentence count f-score is extremely low for the MUHON corpus when compared to the other three corpora. This simple measure provides strong evidence to support the anecdotal perception that operative reports reflect a grammar that is uncommon in the vast majority of written language. Interestingly the discharge report, which also represents text from the clinical record, shows a far higher sentence count f-score.

The purpose of the discharge report is to communicate with other caregivers who will interact with this patient at a later date, whereas the operative report is typically only referred to by the team directly involved in the care of this patient for the current admission. This change in intended audience may be the cause of this change in grammatical structure and formalism of the text. Further analysis would be required to confirm this – including a corpus of post-surgical letters to the referring clinician in the same analysis would provide the most appropriate control in order to isolate the magnitude of the impact of target audience.

It is also probable that the list-like nature of the procedural descriptions contributes to the low accuracy of the sentence segmentation – a very simple update to include hard line breaks in sentence boundary detection algorithms may resolve a large proportion of this issue.

Due to the low reliability of the sentence segmentation within the MUHON corpus, sentence length distribution was not analysed, as it is unlikely to be possible to draw a valid conclusion.

High information density: Again, with respect to information density, the MUHON corpus is seen to be the extreme outlier. Stopword density increases in an unsurprising fashion relative to the level of assumed knowledge in the target audience – from the patient's current direct care team, to future clinicians, to the general public and finally to patients (drug consumers) (MUHON, i2b2, Reference, PIL corpus respectively).

Redundant information is used to facilitate effective transmission of ideas to an unfamiliar audience; however, for the purpose of later self-reference, this is wasteful for both notation and review effort.

Extremely low pronoun and passivity ratios are also indicative of this increase in information density – as the actor and target of each action is often implicit by context (such as an action clearly performed on the patient) or by the action itself (craniotomy acts explicitly on the skull). As such, references to *it*, *that*, *(s)he*, *was* etc. can be dropped without impacting meaning.

This can also be seen in Table 13 – the most distinctive terms between MUHON and the PIL and Reference corpora show a clear lexical difference, whereas the most significant relative frequency differences between MUHON and i2b2 show the corpora to be highly semantically related, but to differ strongly in their pronomial content. The distinctive terms from the i2b2 corpus can all be considered redundant (with the exception of *mg* and *no*) when applying the clipped grammar of the operative report. For example, a typical sentence

She was afebrile, her vital signs were stable in the discharge summary is more likely to be rendered *afebrile; v.s. stable* in the equivalent operative report.

High linguistic diversity: The ratio of tokens/type is significantly lower for the MUHON corpus than any other. This can be interpreted as a high level of linguistic diversity relative to other text types. This is unsurprising based on the high information density within this corpus – as low information words are left out, each retained term is more likely to differ from other included terms. This may also be indicative of a diversity of expression, such as spelling mistakes or abbreviations of the same word, which is less common in more formal text types such as pamphlets or published articles, e.g. *resction/resection*, *polys/polyp*, *hydro/hydration*.

High linguistic diversity is likely to be due also in some part to the combinatorial nature of technical clinical language, where the scientific vocabulary is built of roots, suffixes and prefixes to form very specific terminology (*arthro-*, *bacterio-*, *cardio-*, *-ectomy*, *-graphy*, *-tomy* etc.). The low incidence of negation is also consistent with this characteristic, where a precise term can be formed to indicate a negative,or more technical terms may be selected than the simple list used in this experiment (*non-, an-, absent, atypical*).

The use of specific, scientific terminology can be expected to also affect the corpus of discharge notes to some extent. The i2b2 corpus contains only a middling level of linguistic diversity, however, and therefore the effect can be assumed to be minimal by comparison to the contribution of the staccato grammar employed.

In Figure 10 it can be seen that none of the corpora have sufficient semantic similarity to approach the identity threshold of 0.05, where it can be assumed that the samples are sufficiently similar that they have been created from the same source. The divergence of the MUHON frequency distribution from each of the other included corpora are the three highest – again supporting the case for semantic and syntactic idiosyncrasy, and the need to develop tools which are specifically tuned to this text. MUHON and the Reference corpora have the highest level of divergence, most strikingly seen for very high frequency terms – the slope of the increase between MUHON and PIL is much steeper for high frequency terms, converging to a similar level of divergence as more terms are included. This is likely to be due to the inclusion of common clinical terms in the PIL corpus, so stronger similarity is seen for general/high level terms such as *pain*, *blood*, *joint* etc., but this is dwarfed by the significant difference in structure and semantic content once lower frequency terms are included.

3.5.3 Operative report sublanguage: conclusion

The results seen here indicate that the purpose and likely the intended audience have a stronger effect on observed linguistic features than the content or clinical nature of text. This therefore refutes the null hypothesis, and implies that NLP applications developed for the general clinical record will perform poorly when applied without refinement on operative reports.

It would be valuable to collate a fifth corpus, containing post-operative letters to referring clinicians. By providing strictly semantically matched content for operative reports, it would be possible to confirm or contradict the hypothesis that the intended audience for later reference is a key differentiator in the structure and composition of clinical text.

There are many subtasks that must be tackled in order to move forward the development/tuning of NLP tools to deal with operative reports. Based on the results seen here, high priority should be given to resolving the unreliable performance of general purpose sentence segmenting algorithms, thereby providing a solid basis for syntactic parsing. Once this is achieved, domain specific lexical enrichment (both abbreviations and idiosyncratic terms) is likely to be a worthwhile next step, given the high linguistic diversity observed. The development of a large tagged corpus is also required in order to allow unsupervised algorithms to perform reliable information extraction under the clipped grammar and high level of assumed/implicit knowledge that is present in operative reports.

Table 12: K-L divergence of term probability distributions

n terms	PIL/i2b2	MUHON/PIL	PIL/Ref	MUHON/i2b2	i2b2/Ref	MUHON/Ref
100	1.624740213	2.66765031	1.742293249	2.063121931	1.999214561	2.789538474
200	1.567467605	2.35204421	1.696258545	1.902352466	1.882862264	2.540080473
300	1.388122954	2.212521445	1.623366061	1.786785791	1.776283298	2.423724863
400	1.337323381	2.115531229	1.550393982	1.727642744	1.715199016	2.323053807
500	1.288743599	1.993618698	1.489903635	1.666077109	1.659913819	2.230949862
1,000	1.155648675	1.783619874	1.337485035	1.47932908	1.482772805	1.963735389
1,500	1.09718094	1.683349063	1.267883359	1.399923628	1.404990601	1.845728219

Table 13: Log likelihood – 10 most distinctive terms relative to MUHON for each experimental corpus

MUHON	PIL Corpus	LL	MUHON	i2b2	LL	MUHON	Ref	LL
	.	4806.86		the	47020.00		the	1145.59
	the	4620.92	and		34420.29		a	456.21
	you	3750.92		was	32208.39	left		352.87
	your	3233.33	of		27748.00	right		244.82
	or	2173.08	to		24077.69	of		206.90
	of	1926.47		a	21585.44	vicryl		191.93
	to	1903.52	with		18246.12	analgesia		183.58
	a	1745.01		patient	18175.01	monocryl		179.41
	if	1614.18	on		16518.64	in		175.90
	is	1571.25	in		13403.14	routine		167.66
and		824.12	for		11542.65		is	167.33
with		362.12		mg	10596.35		said	166.51
on		351.98		he	10004.37	to		160.58
vicryl		202.06		she	9972.85		it	142.45
analgesia		193.27		no	8651.48	knee		133.51
monocryl		188.88		is	7862.18		million	101.55
left		136.56		his	7641.64		mr.	99.43
when		134.65	discharge		7474.59		are	97.84
knee		131.84	at		7255.97		was	97.31
excision		131.78	as		4991.06		by	93.23

4 Text mining of operative reports

The work presented in Chapter 2 provides a foundational understanding of the types of concepts which are available in a typical operative report (see Table 6). The gap analysis performed has demonstrated that much of this information is unavailable from standard billing codes alone, and thus significant value is gained by the encoding of these reports into a more appropriate knowledge model for the purpose of automated analysis, monitoring and surveillance – SNOMED CT was established as the optimal knowledge model in this domain. This chapter will therefore demonstrate initial work to create an automated encoding system for SNOMED CT concepts, based on the statistical analysis of the text of operative reports as discussed in Chapter 3.

4.1 Text mining of operative reports: methods

4.1.1 Target definition

A set of target concepts have been selected as proof of concept. Identification of the full SNOMED CT representation is beyond the scope of both this project and the small set of reports available for analysis, and is left for future work. The selection of tasks was prioritised for proof of concept development based on:

1. With the chosen target structure in mind, which of the components are best suited for the purposes of automated extraction?
2. What information is available within the written reports that is not available unambiguously in the billing codes?
3. Out of this available information, what is most clinically relevant for the purposes of automated monitoring of patient safety?

Chute et al. [51] note that SNOMED CT outperforms other observed encoding systems in all concept categories, but that this is particularly noteworthy for modifiers, which exhibit extremely low levels of coverage elsewhere. From Table 6 it can be seen that locative modifiers, which exhibit some of the lowest levels of coverage for identified in operative reports also fall under this classification [14].

The target for this extraction is therefore defined as all concepts falling under the SNOMED CT term *309825002: Spatial and relational concepts (qualifier value)*. There are 1110 children concepts, covering procedural approach, relative sites and surgical access values. These qualifiers are useful as input for a number of different types of automated patient safety interventions. The *approach* qualifiers can form part of a relative risk calculation where

[14]Rejecting *Devices* and *Findings* from analysis due to their likelihood to be available elsewhere in the EHR, and *Anaesthesia*, *Positioning*, *Purpose* and *Tool* for their overall low frequencies.

a different surgical approach is indicative of underlying patient characteristics and are also required for comparative effectiveness measure where techniques and their associated outcomes are evaluated. *Site* and *relative location* concepts are required for follow-up measures where surgical revisions can be correctly matched to their initial procedures. Any of the qualifier values may also be used for observational studies or as a measure of clinicians' propensity to adhere to guidelines.

4.1.2 Sentence boundary detection

It is first necessary to correct the very low performance sentence segmentation as observed in Chapter 3. This is not a particularly complex task compared to other text mining activities; however, it is a foundational step in the NLP pipeline – errors will propagate to later phases, where concepts may be missed which break across false-positive sentence boundaries, or relationships may be wrongly resolved when a sentence break is missed. Consider the abbreviation *neuro. obvs.* (neurological observations) which appears 7 times in the MUHON corpus, both with (3) and without (4) punctuation. With some domain knowledge, it is clear to a human reader that the first period does not indicate a true sentence break as it leaves an adjective without an associated object; however, it is unlikely to be available within a list of known abbreviations created over standard English text, and thus will be missed by the majority of standard sentence boundary detection algorithms.

By observation, the primary reasons for the extremely low accuracy of standard sentence segmentation on the operative reports are (1) inconsistent use of standard capitalisation, which is likely to be encoded as an indicator for sentence breaks and (2) the prevalence of short, ungrammatical sentences demarcated by newlines alone.

A freely available support vector machine (SVM) classifier [88] was trained using 10% of available data. In the evaluation data, 1202 sentences were detected correctly, with 28 false positives and no false negatives, giving a precision of 98.85%, recall of 100% and f-score of 97.72%. This far outperforms the naïve classifier (f-score 19.05%), and approaches the performance when applied to a more standard corpus (f-score 99.71%). The remaining false positives fall into three categories:

1. The use of a question mark at the beginning of a sentence to indicate a query (*? discharge today*)
2. Enumerated lists in the middle of otherwise normally formatted paragraphs (*1. mobilise . . . 2. drain . . .*)
3. Rare abbreviations which were not identified from training data (*subcut. maxolon*)

The first two of these conditions are very unlikely to affect named entity recognition tasks – despite forming false-positive sentence boundaries, no concepts are broken across them. It is however possible that these will be important contextual clues for downstream text processing activities such as detecting the level of certainty of a statement, and temporal resolution respectively.

Rare abbreviations (even varying on a per-clinician basis) are likely to cause more

issues in practice. Even extremely large corpus sizes will struggle to predict truly personal abbreviations and therefore a more robust way of handling them must be identified.

4.2 Baseline

4.2.1 Baseline: methods

If an algorithm does not improve the accuracy of a given task above the most obvious or base method, then its value is limited to interest only. Therefore, for the purposes of baseline, direct dictionary matching and two popular existing tools were applied to the reports.

For these comparisons, if a term was present more than once in the text in the same context (i.e. a modifier referring to the same element) this was counted as a single match or miss as appropriate.

After data cleaning, 881 of 901 reports were included. Excluded reports did not have any text in the *Operation Performed*, *Details of Operation* or *Operation Findings* fields.

All reports were annotated with their surgical specialty, location, surgical approach, sidedness and whether they were an open or closed procedure. Only those elements which could be successfully coded using the SNOMED CT sub-hierarchy identified above were annotated. *Unknown* and *uncodable* qualifiers were therefore also included.

Direct text matching: A straightforward dictionary matching algorithm was developed which considered terms in their raw form.

All possible n-grams were created for each sentence in each report, up to n=7 (reflecting the longest SNOMED CT term under consideration). As mentioned in the previous section, this included a small number of false positive sentence boundaries, however this was tolerated for the purpose of devising the simplest possible text matching method.

These n-grams were matched directly against the preferred terms of the subset of SNOMED CT of interest. In the instance of a term being included in more than one possible match, the longest available match was returned. A simple thesaurus was also constructed which included all UMLS synonym terms and matched in the same manner, for the sake of comparability against existing tools.

Existing tools: The existing tools MetaMap [89] and cTAKES [69] were applied in order to assess performance of widely used, freely available clinical information extraction tools against the algorithms developed here. These tools were chosen as they far outrank other known tools by number of published PubMed articles in the last 5 years and therefore are assumed to be the most commonly implemented.

The optional inbuilt word sense disambiguation functionality of MetaMap was used in order to filter results. The `AggregatePlaintextFastUMLSProcessor` analysis engine was used to run cTAKES.

Table 14: Accuracy of baseline information extraction techniques

	Match Count	Correct Matches (%)	False Matches (%)
MetaMap	5,972	4,821 (80.1)	1,151 (19.9)
cTAKES	6,845	6,129 (89.5)	716 (10.5)
Direct text matching	819	792 (96.7)	27 (3.3)

4.2.2 Baseline: results and discussion

Table 14 shows the differences between the number of matches which are made when processing the raw input text under the three strategies defined above.

MetaMap is greedy – matching all possible non-overlapping terms, irrespective of context, and then using a scoring algorithm based on configurable settings, together with the input text, to determine the statistically most likely match out of these candidate matches [90]. This results in a somewhat higher false positive rate than either cTAKES or straight dictionary matching.

cTAKES takes a more comprehensive NLP approach [69]. Where MetaMap's preprocessing consists of only phrase segmentation, cTAKES performs sentence boundary detection, tokenization, normalization, part-of-speech (POS) tagging and shallow parsing before attempting named entity recognition. It is able to create more matches than MetaMap over the same text while relying on the same underlying UMLS synonym dictionary by a combination of custom lexical enrichment and by not imposing the same restriction for non-overlapping terms.

Unsurprisingly, the simple dictionary algorithm produces far fewer matches than either of the other tools, even with the inclusion of UMLS thesaurus terms. From this, it can be seen that despite the differences identified in Chapter 3 between general clinical text and the operative report, there is sufficient similarity for these tools to perform well for certain tasks. cTAKES in particular has been developed with an architecture geared for extension and customisation and shows potential for being adapted to the operative report.

Table 15 contains a breakdown of results where these tools were applied to four specific subtasks: identifying surgical site, approach, sidedness and whether the procedure was open or closed. cTAKES generally outperforms the other two tools when applied to these targeted tasks, except in the instance of determining the side of the body on which a procedure was performed. This task is trivial compared to some of the complex concepts which can be recognised by this application; however, it clearly does not form part of the cTAKES knowledge base.

Table 15: Baseline results

	Site	Approach	Sidedness	Open/Closed
		Metamap baseline		
Precision	0.909	0.869	0.932	0.930
Recall	0.125	0.112	0.925	0.308
f-score	0.220	0.198	0.929	0.462
		cTAKES baseline		
Precision	0.994	0.992	0	0.988
Recall	0.443	0.750	0	0.676
f-score	0.613	0.854	N/A	0.796
		Dictionary match baseline		
Precision	0.921	0.913	0.992	0.911
Recall	0.097	0.131	1.00	0.106
f-score	0.176	0.230	0.996	0.191

4.3 Classifier development

Given the poor performance of simple term-matching seen above, this next set of experiments was designed to determine the feasibility of using statistical text classification to automatically identify concepts within text from operative reports. The tasks of identifying sidedness (left/right/bilateral) and openness of a given procedure, if done effectively, will allow a good automated matching to concepts from the SNOMED CT sub-hierarchy *Spatial and relational concepts*.

Site of operation and surgical approach classifiers were not built – given the large number of potential matches relative to the corpus size this was not likely to be successful. Instead, automated identification of surgical specialty was attempted in order to provide a meaningful filter of identified concepts that can reduce the high level of false positives seen in the baseline methods.

4.3.1 Classifier: methods

Pre-processing: The 881 reports (plus their annotations as described above) were first pre-processed to allow effective feature extraction. All punctuation and non-alphanumeric text was removed from text, and numbers were replaced with a placeholder (NUM). Text case was normalised and stop words removed.

Feature extraction: The operative reports were used to create a bag-of-words frequency distribution as the most basic set of features. Bi-gram, tri-gram and lemmatized term features

were also created.

A list of medical suffixes and affixes was collated to create a clinically tuned stemming algorithm *MedStem* that is able to stem for clinically relevant matches at either the beginning or the end of a word. If a word matched both a suffix and an affix in the source list, both features were counted (e.g. *arthroplasty* matches both *arthro-* – of or relating to joints – and *-plasty* – repair or reconstruction).

In addition, a simple spellchecking algorithm was developed. Words contained within the source SNOMED terms were used as the dictionary of known words. If a word in the input text was not in the dictionary of known words, common spelling errors (missing, transposed, replaced or inserted letters) were derived and checked against this same dictionary. If more than one alternate spelling was present in the known words, the most likely spelling was returned (by dictionary frequency).

Classifier design: A Support Vector Machine (SVM) classifier was chosen as it was expected to be effective in the high-dimensional space of surgical specialty. SVMs are however, sensitive to class imbalance [91] and therefore classifier training was performed both with the original data set and with a data set that had been balanced by undersampling.

L2 regularization was applied in order to compensate for the small number of training samples and large number of parameters used as input and avoid subsequent overfitting.

Classification: A 10-fold cross-validation methodology was followed, whereby sets of training (60%), validation (20%) and testing (20%) data were randomly assigned 10 times, in order to ensure generalisability to new data. The average performance over 10 trials is reported here.

4.3.2 Classifier: results and discussion

Precision, recall and f-score were calculated for each feature set when applied to each task. In the case where the target domain was not binary true/false, the scores were averaged across all target labels.

The results of the text classification are shown in Table 17. Overall, it can be seen that for well-defined tasks, text classifiers outperform existing tools. Whether a surgery is open or closed can be detected with high precision (0.945) and an f-score of 0.931, compared to the best performing available method (cTAKES), which had an f-score of 0.796. Whilst cTAKES has acceptable precision for this task, the poor recall reflects the significant differences that have been identified between the operative report and standard clinical text – a high number of strings cannot be recognised due to their atypical expression, as expected. In future work, an operative report-specific lexicon, including the detection of abbreviations and acronyms in common usage, should be collated and integrated with the existing cTAKES knowledge base to achieve improved performance.

Although the detection of surgical specialty is not directly comparable with the baseline site and approach measures, the high performance of this measure on a balanced training set indicates the potential for developing knowledge-based system which can be used to simultaneously relax matching rules (e.g. to allow partial or misspelled matches) without increasing false positive matches by applying smarter filters to reject extraneous matches.

Each task clearly has a different profile for optimal classifier design. This is an important finding, as it shows that a single approach is not going to be applicable to all problems, instead an iterative workflow of classification and validation is required. This workflow is expensive, as manual labelling and tuning is required for each sub-step; however, it seems unavoidable based on the disparities seen here.

Feature set size (see Table 16) was reduced by at most 25%, by the application of a lemmatizing algorithm. Bigram and trigram classifiers increased feature set size, however for the sake of computational efficiency and overall classifier performance, all but the 2,609 most frequent features (equivalent in size to the raw feature set) were discarded.

Balancing the training set is an effective way of improving classifier performance in most instances, however its effect is much more significant for the tasks of identifying surgical specialty and laterality than for identifying whether a surgical approach is open or closed. Surgical specialty labels, however, have the highest number of target classes, and are therefore more affected by the undersampling technique – losing more than 80% of the training set and having as low as 10 samples corresponding to each label. This shows that the classifiers are somewhat resilient to the low training set size; however, further work must be done in order to determine the learning curve of each classifier and therefore the point at which increasing the labelling effort gives only diminishing returns.

It is intuitive that the *MedStem* routine is the most effective for determining whether a procedure is open or closed (f-score 0.931) due to the way in which clinical terms are built – laparo*scopy*, cysto*scopy* and arthro*scopy* being grouped together based on their suffix creates a much more strongly indicative feature than any of these terms individually.

Similarly, the high performance of bigram and trigram features makes sense when reviewing the most informative bigram and trigram features – many of these contain both an action and a location, which aligns with expected surgical specialty strongly differentiating features.

The task of identifying the side (left/right/bilateral) of a surgical procedure seems at a surface level to be the most trivial, however, this is seen to perform the most erratically in practice. This is likely to be due to the low number of meaningful features which are then overwhelmed with erroneous features in this small sample size. As seen in the previous section (Table 15), this task is better suited to an heuristic search method and performs very well without the assistance of a classifier (f-score 0.996).

4.3.3 Classifier: conclusion

There is no one-size-fits-all approach to NLP for operative reports. These experiments have demonstrated that for each encoding sub-task, an individually tuned approach is required, which may be either heuristic or statistical. This is a high effort and high cost solution, however given the restricted input domain, can be expected to pay off with high accuracy classification.

These experiments are limited by the relatively small sample size that was made available for development, however with a balanced training strategy it is still possible to achieve good results. Additional data should be collected in order to understand the learning

Table 16: Feature set size

	Feature set size	%
Raw input	2,609	100
Spell-checked	2,116	81
Lemmatized	1,956	75
MedStemmed	2,142	82
Bigrams	6,721	258
Trigrams	8,035	308

Table 17: Text classifier results

	Open/Closed			Surgical Specialty			Sidedness		
	Precision	**Sensitivity**	**F1-Score**	**Precision**	**Sensitivity**	**F1-Score**	**Precision**	**Sensitivity**	**F1-Score**
				Raw					
Whole set	0.931	0.905	0.918	0.808	0.658	0.725	0.636	0.515	0.569
Balanced	0.937	0.901	0.918	0.708	0.990	0.826	0.873	0.929	0.900
				Spelled					
Whole set	0.926	0.906	0.916	0.809	0.658	0.726	0.524	0.511	0.518
Balanced	0.933	0.892	0.912	0.710	0.985	0.825	0.810	0.926	0.864
				Lemmatized					
Whole set	0.899	0.890	0.894	0.804	0.663	0.727	0.545	0.496	0.519
Balanced	0.919	0.854	0.885	0.719	0.952	0.819	0.908	0.945	0.926
				Med-Stemmed					
Whole set	0.927	0.925	0.926	0.754	0.630	0.687	0.520	0.499	0.509
Balanced	0.945	0.919	0.931	0.759	0.984	0.857	0.431	0.935	0.590
				Bigrams					
Whole set	0.908	0.802	0.852	0.826	0.544	0.656	0.730	0.584	0.649
Balanced	0.897	0.777	0.832	0.882	1.000	0.937	0.788	0.925	0.851
				Trigrams					
Whole set	0.865	0.737	0.796	0.757	0.475	0.584	0.625	0.479	0.542
Balanced	0.875	0.654	0.749	0.933	1.000	0.965	0.764	0.845	0.802

curve of these classifiers and thus estimate a safe minimum training set size before producing any production classifiers.

While it has been demonstrated that significant differences exist between the text in operative reports and general clinical text, it is seen that tools developed for general clinical text still perform well for certain tasks. cTAKES in particular, due to its extensible architecture, has been identified as a good candidate upon which to base future work. It is possible to extend the cTAKES knowledge base and analytical engines for new textual sources and output tasks.

Based on the work done here, it is likely that the most productive next step towards a fully automated SNOMED CT encoding of operative reports is to curate an operative report-specific synonym knowledge base. Heuristic detection of abbreviations and acronyms which can then be manually defined and implemented in the cTAKES processing pipeline is expected to have significant positive impact results seen above. Given the diversity of language seen in these reports, a large number of additional reports will be required for this step.

5 Conclusion

The work presented in this thesis is intended to provide a broad foundation upon which to base efforts toward automated surveillance of patient safety in the surgical domain. This chapter summarises the findings and then presents areas identified for future research.

5.1 Summary of findings

5.1.1 Limitations of the current practice of surgery characterisation

For the purposes of automated patient safety interventions and observational studies, the current practice of characterisation of surgeries within a typical EHR is inadequate. A gap analysis was performed on a set of operative reports and it was found that only 31% of the data contained in operative reports is available unambiguously within the billing codes.

The data that is available in the written report but cannot be retrieved from the billing codes includes clinically relevant elements such as surgical approach, relative location and peri-operative observations. These elements are valuable for automated surveillance and comparative effectiveness measures, both of which are lacking in current surgical practice, which relies instead on manual reporting and case studies. Automated encoding will create opportunities for significantly improved electronic reporting of surgical adverse events – a process which is currently inconsistently and manually managed.

To be able to release this data for automated analysis, it is necessary to identify an appropriate choice of target knowledge model that can effectively and completely contain this information. Through a review of ontologies that are in use to represent surgeries for many different purposes, SNOMED CT was identified as the best candidate model due to its flexibility, granularity, wide adoption and hierarchical nature. Manual encoding of a test set of operative reports into SNOMED CT demonstrated 95% unambiguous coverage of clinically relevant concepts.

5.1.2 Empirical analysis of the text of operative reports

A review of natural language processing systems in the clinical domain was undertaken and showed no existing tools that were built with operative reports forming part of either the input corpus or test set. Based on this, statistical textual analysis was performed in order to estimate the applicability of general clinical tools to the sublanguage of operative reports.

The sublanguage of operative reports was found to be not representative of the language of general clinical text. It is characterised by extremely high information density and linguistic diversity, with redundant information omitted – including pronouns and even the subject and/or object of certain verbs, where these are implicit in the narrow context of the report. There is also significant variability of expression (grammar, spelling, abbreviation),

which can be tolerated by the target audience of the report (the immediate patient care team) as opposed to more formal components of the clinical record.

This lack of coherence implies that tools developed for the EHR as a whole, or other specific subsets of clinical language (such as radiology reports) cannot be assumed to be directly applicable to the operative report.

5.1.3 Automation of encoding: proof of concept

A set of classifiers were developed that allow the automated encoding of a subset of the target ontology. These classifiers achieved f-scores of 0.945 and 0.965 for the surgical approach (open or closed) and surgical specialty respectively. The laterality of a procedure was more effectively detected with a simple heuristic approach (f-score 0.996).

The experimental corpus available for this project was limited, however it was still possible to build text classifiers with reasonable accuracy for certain well-defined classification tasks. This demonstrates the coherence of the text and the general feasibility of classifier design in the pursuit of the goal of automated encoding of surgical procedures.

Each classification task had a distinctly different optimal design profile, which suggests that an iterative design approach is required.

Despite the distinctive sublanguage which was observed in operative reports, the existing clinical NLP tool cTAKES was found to perform relatively well for some tasks (e.g. f-score 0.854 for detection of surgical approach), and to have a low number of false positive matches across the board. The high number of false negatives, by comparison, is consistent with the observed informal spelling and casual style of abbreviation in operative reports. Operative report-specific lexical enrichment is expected to go a long way towards resolving this issue. There were certainly some gaps in its ability (e.g. precision and recall were both zero for the detection of laterality of surgical procedure); however, the extensible nature of the knowledge base and analytical engine implies that cTAKES may still provide a strong foundation for further work.

5.2 Areas of future research

A wider project encompassing the following areas of research is planned to immediately follow this work:

5.2.1 Corpus expansion

The corpus used in this work was restricted by coincidental EHR upgrades at MUH, reducing the availability of personnel to provide the requested volume of reports. Now that this upgrade effort has concluded, the opportunity to collate a much larger data set is available.

With an expanded corpus, a key initial activity is to investigate the learning curve of these classifiers in order to determine a sensible and practically achievable corpus size that

is necessarily to reliably build additional successful classifiers.

In addition, both theoretical and practical observations (Chapters 3 and 4 respectively) have shown that lexical enrichment with operative report definitions, synonyms and abbreviations is certainly the single most necessary step to improve the performance of NLP tools when applied to operative reports.

5.2.2 Application of SNOMED CT encoding to real-world problems

SNOMED CT encoding was chosen in part due to the deep knowledge that is provided by its hierarchy. The detection of surgical specialty is the first step towards a knowledge-based NLP model, where filters and named entity recognition can become 'smarter' based not only on the immediate surrounding text but also on the broader characteristics of the report and procedure. The application of these classifiers to provide contextual clues for encoding must be thoroughly explored.

This project has identified a theoretical best representation of surgical procedures for the automation of surveillance, quality and comparative effectiveness measures in SNOMED CT; however, validation of this selection in practice is still required. The successful design, implementation and evaluation of an automated platform for one or more of these goals would provide unequivocal evidence for this selection.

In order to achieve this, it will be valuable to expand the scope of interviews with the quality and safety teams on the front-line at MUH to break down the incident reporting and investigation workflow into its component steps. From this, targets for further classifier development and other encoding subtasks can be completed and evaluated in practice. These targets will be amalgamated into a single tool which can aggregate hospital and department quality and safety performance.

5.3 Conclusion

Operative reports contain valuable information which is currently far from being fully exploited in the pursuit of improving patient safety. Billing codes provide only a small proportion of this information in a format that is available for surveillance and research, and therefore a system which can automatically encode a higher proportion of the written report into structured information is required.

SNOMED CT is believed to be the best available structured format for these purposes, and this project has demonstrated that it is feasible to use a combination of NLP techniques to perform this encoding, although modifications are required if existing clinical tools are to be used.

References

[1] World Health Organization (WHO), "WHO draft guidelines for adverse event reporting and learning systems: from information to action," www.osp.od.nih.gov/sites/default/files/resources/Reporting_Guidelines.pdf, 2005, accessed: 2015-04-09.

[2] World Health Organization (WHO), "Toolkit on monitoring health systems strengthening: service delivery DRAFT," www.who.int/healthinfo/statistics/toolkit_hss/EN_PDF_Toolkit_HSS_ServiceDelivery.pdf, 2008, accessed: 2015-04-09.

[3] National Institute for Health and Clinical Excellence (NICE), "Interventional procedures programme process guide," www.nice.org.uk/media/default/about/what-we-do/nice-guidance/nice-interventional-procedures/interventional-procedures-programme-process-guide.pdf, 2009, accessed: 2015-04-09.

[4] T. G. Weiser, S. E. Regenbogen, K. D. Thompson, A. B. Haynes, S. R. Lipsitz *et al.*, "An estimation of the global volume of surgery: a modelling strategy based on available data," *The Lancet*, vol. 372, no. 9633, pp. 139–144, 2008.

[5] World Health Organization (WHO), "Conceptual framework for the international classification for patient safety: Version 1.1," www.who.int/patientsafety/taxonomy/icps_full_report.pdf, 2009, accessed: 2015-06-15.

[6] World Health Organization (WHO), "The international classification for patient safety – general description," "www.who.int/patientsafety/implementation/taxonomy/icps_general_description.pdf", 2009, accessed: 2015-07-01.

[7] NSW Department of Health, "Health administration act no 135 section 20l," Parliamentary Counsel's Office, 1982.

[8] NSW Department of Health, "Incident management policy," http://www0.health.nsw.gov.au/policies/pd/2014/pdf/PD2014_004.pdf, pp. 40–41, 2014, accessed: 2015-07-21.

[9] Clinical Excellence Commission, "Collaborating Hospitals' Audit of Surgical Mortality," http://www.cec.health.nsw.gov.au/programs/chasm, 2015, accessed: 2015-06-19.

[10] International Conference on Harmonisation, "Clinical safety data management: Data elements for transmission of individual case safety reports – E2B (R2)," www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2B/Step4/E2B_R2__Guideline.pdf, 2001, accessed: 2015-04-16.

[11] U.S. Department of Health and Human Services – Food and Drug Administration, "Guidance for industry: Good pharmacovigilance practices and pharmacoepidemiologic assessment," www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf, 2005, accessed: 2015-05-04.

[12] World Health Organisation (WHO), "The importance of pharmacovigilance," www.apps.who.int/iris/bitstream/10665/42493/1/a75646.pdf, 2002, accessed: 2015-04-09.

[13] M. Inacio, E. W. Paxton, Y. Chen, J. Harris, E. Eck *et al.*, "Leveraging electronic medical records for surveillance of surgical site infection in a total joint replacement population," *Infection Control*, vol. 32, no. 04, pp. 351–359, 2011.

[14] Collaborating Hospitals' Audit of Surgical Mortality, "Surgical case form: Version 2.0," www.cec.health.nsw.gov.au/__data/assets/pdf_file/0019/260902/CHASM-SCF-template-version-2.0.pdf, accessed: 2015-03-18.

[15] M. S. Calderwood, A. Ma, Y. M. Khan, M. A. Olsen, D. W. Bratzler *et al.*, "Use of medicare diagnosis and procedure codes to improve detection of surgical site infections following hip arthroplasty, knee arthroplasty, and vascular surgery," *Infection Control*, vol. 33, no. 01, pp. 40–49, 2012.

[16] P. McCullouch, D. G. Altman, W. B. Campbell, D. R. Flum, P. Glasziou *et al.*, "No surgical innovation without evaluation: the IDEAL recommendations," *The Lancet*, vol. 374, no. 9695, pp. 1105–1112, 2009.

[17] The Lancet, "Surgical research: the reality and the IDEAL," *The Lancet*, vol. 374, no. 9695, p. 1037, 2009.

[18] R. Horton, "Surgical research or comic opera: questions, but few answers," *The Lancet*, vol. 347, no. 9007, pp. 984–985, 1996.

[19] E. Coeira, *Guide to health informatics (Third Edition).* CRC Press, 2015, pp. 33–61.

[20] R. Studer, V. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data & Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[21] M. Ivanović and Z. Budimac, "An overview of ontologies and data resources in medical domains," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5158–5166, 2014.

[22] N. Gaurino, *Formal Ontology and Information Systems.* Amsterdam: IOS Press, 1998, pp. 3–15.

[23] P. Spyns, R. Meersman, and M. Jarrar, "Data modelling versus ontology engineering," *ACM SIGMOD Record*, vol. 31, no. 4, pp. 12–17, 2002.

[24] A. Gómez Pérez, "Evaluation of ontologies," *International Journal of Intelligent Systems*, vol. 16, no. 3, pp. 391–409, 2001.

[25] D. Vrandečić, "Ontology evaluation," in *Handbook on Ontologies (Second Edition)*, S. Staab and R. Studer, Eds. Springer Berlin Heidelberg, 2009, pp. 293–313.

[26] A. Gómez Pérez, "Ontology evaluation," in *Handbook on Ontologies in Information Systems (First Edition)*, S. Staab and R. Studer, Eds. Springer Berlin Heidelberg, 2004, pp. 251–274.

[27] L. Orbst, W. Ceusters, I. Mani, S. Ray, and B. Smith, *The Evaluation of Ontologies.* Springer US, 2007, pp. 139–158.

[28] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International Journal of Human-Computer Studies*, vol. 43, pp. 907–928, 1995.

[29] H. Tan and P. Lambrix, "Selecting an ontology for biomedical text mining," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Association for Computational Linguistics.*, 1999, Conference Proceedings, pp. 55–62.

[30] J. Baumeister and D. Seipel, "Smelly OWLs – design anomalies in ontologies." in *FLAIRS Conference*, vol. 215, 2005, Conference Proceedings, p. 220.

[31] M. Fahad, M. A. Qadir, and M. W. Noshairwan, "Semantic inconsistency errors in ontology," in *IEEE International Conference on Granular Computing, 2007. GRC 2007.* IEEE, 2007, Conference Proceedings, pp. 283–283.

[32] T. Bittner, M. Donnelly, and S. Winter, "Ontology and semantic interoperability." in *Large-scale 3D data integration: Challenges and Opportunities.* CRC Press, 2005, pp. 139–160.

[33] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *Computational Intelligence Magazine, IEEE*, vol. 9, no. 2, pp. 48–57, 2014.

[34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.*, "Gene ontology: tool for the unification of biology," *Nature*, vol. 25, pp. 25–29, 2000.

[35] M. Rospocher and L. Serafini, "An ontological framework for decision support," in *Semantic Technology.* Springer Berlin Heidelberg, 2013, pp. 239–254.

[36] D. Wang, M. Peleg, S. Tu, A. Boxwala, R. Greenes *et al.*, "Representation primitives, process models and patient data in computer-interpretable clinical practice guidelines: A literature review of guideline representation models," *International journal of medical informatics*, vol. 68, no. 1, pp. 59–70, 2002.

[37] T. Oliveira, P. Novais, and J. Neves, "Assessing an ontology for the representation of clinical protocols in decision support systems," in *Trends in Practical Applications of Agents, Multi-Agent Systems and Sustainability.* Springer, 2015, pp. 47–54.

[38] W. Runciman, Personal Interview, July 13th 2015.

[39] B. Czerniec, Personal Interview, June 26th 2015.

[40] U.S. National Library of Medicine, "Supporting interoperability - terminology, subsets and other resources from NLM," http://www.nlm.nih.gov/hit_interoperability.html, 2013, accessed: 2015-08-24.

[41] B. Trombert-Paviot, J. Rodrigues, J. Rogers, R. Baud, E. van der Haring *et al.*, "GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures," *International Journal of Medical Informatics*, vol. 58-59, pp. 71–85, 2000.

[42] OpenGALEN, "The OpenGALEN manifesto," www.opengalen.org/manifest/intro.html, 2015, accessed: 2015-06-19.

[43] J. Rodrigues, A. Rector, P. Zanstra, R. Baud, K. Innes *et al.*, "An ontology driven collaborative development for biomedical terminologies: from the French CCAM to the Australian ICHI coding system," *Studies in health technology and informatics*, vol. 124, pp. 863–868, 2005.

[44] B. Trombert-Paviot, A. Rector, R. Baud, P. Zanstra, C. Martin *et al.*, "The development of CCAM: the new French coding system of clinical procedures," *Health Information Management Journal*, vol. 31, no. 1, pp. 1–11, 2003.

[45] A. Rector, A. Rossi, M. Consorti, and P. Zanstra, "Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN organisation," *International Journal of Medical Informatics*, vol. 48, pp. 71–84, 1998.

[46] A. Rector, W. Nowlan, and A. Glowinski, "Goals for concept representation in the GALEN project," in *Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association*, 1993, Conference Proceedings.

[47] F. Lalys and P. Jannin, "Surgical process modelling: a review," *International Journal of Computer Assisted Radiology*, vol. 9, no. 3, pp. 495–511, 2014.

[48] N. Padoy, T. Blum, S. Ahmadi, H. Feussner, M. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, vol. 16, no. 3, pp. 632–641, 2012.

[49] D. Katić, A.-L. Wekerle, F. Gärtner, H. Kenngott, B. P. Müller-Stich *et al.*, "Ontology-based prediction of surgical events in laparoscopic surgery," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2013, pp. 86 711A–86 711A.

[50] J. Rodrigues, B. Trombert-Paviot, R. Baud, J. Wagner, and F. Meusnier-Carriot, "GALEN-In-Use: Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures," *Studies in health technology and informatics*, vol. 1, pp. 623–627, 1998.

[51] C. G. Chute, S. P. Cohn, K. E. Campbell, D. E. Oliver, J. R. Campbell *et al.*, "The content coverage of clinical classifications," *Journal of the American Medical Informatics Association*, vol. 3, pp. 224–233, 1996.

[52] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen, "Ontology-based error detection in SNOMED-CT," in *Proceedings of MEDINFO*, vol. 2004, 2004, Conference Proceedings, pp. 482–6.

[53] W. Ceusters, B. Smith, and J. Flanagan, "Ontology and medical terminology: Why description logics are not enough," in *Towards an Electronic Patient Record (TEPR 2003), Boston, MA*, 2003, Conference Proceedings.

[54] A. L. Rector, S. Brandt, and T. Schneider, "Getting the foot out of the pelvis: modeling problems affecting use of snomed ct hierarchies in practical applications," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 432–440, 2011.

[55] J. E. Rogers, C. Price, A. Rector, W. D. Solomon, and N. Smejko, "Validating clinical terminology structures: integration and cross-validation of Read Thesaurus and GALEN." in *Proceedings of the AMIA Symposium, American Medical Informatics Association*, 1998, Conference Proceedings, p. 845.

[56] A. Rector and J. Rogers, "Patterns, properties and minimizing commitment: Reconstruction of the GALEN upper ontology in OWL," in *Proceedings of the EKAW*, vol. 4, 2004, Conference Proceedings.

[57] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[58] G. Heja, G. Surján, G. Lukácsy, P. Pallinger, and M. Gergely, "GALEN based formal representation of ICD10," *International journal of medical informatics*, vol. 76, no. 2, pp. 118–123, 2007.

[59] C. M. DesRoches, D. Charles, M. F. Furukawa, M. S. Joshi, P. Kralovec *et al.*, "Adoption of electronic health records grows rapidly, but fewer than half of us hospitals had at least a basic system in 2012," *Health Affairs*, pp. 10–1377, 2013.

[60] J. S. McCullough, "The adoption of hospital information systems," *Health Economics*, vol. 17, no. 5, pp. 649–664, 2008.

[61] G. Hripcsak, S. Bakken, P. D. Stetson, and V. L. Patel, "Mining complex clinical data for patient safety research: a framework for event discovery," *Journal of Biomedical Informatics*, vol. 36, no. 1-2, pp. 120–130, 2003.

[62] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, p. 142, 1994.

[63] N. Sager, C. Friedman, and M. S. Lyman, *Medical language processing: computer management of narrative data*. Addison-Wesley, 1987.

[64] C. Friedman, "A broad-coverage natural language processing system," in *Proceedings of the AMIA Symposium, American Medical Informatics Association*, 2000, Conference Proceedings.

[65] C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994.

[66] P. Haug, S. Koehler, L. M. Lau, P. Wang, R. Rocha, and S. Huff, "A natural language understanding system combining syntactic and semantic techniques." in *Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association*, 1994, Conference Proceedings, p. 247.

[67] P. Haug, S. Koehler, L. Lau, P. Wang, R. Rocha, and S. Huff, "Experience with a mixed semantic/syntactic parser," in *Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association*, 1995, Conference Proceedings.

[68] S. Koehler, "Symtext: a natural language understanding system for encoding free text medical data," Thesis, The University of Utah, 1998.

[69] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[70] Q. T. Zeng, S. Goryachev, S. Weiss, M. Sordo, S. N. Murphy, and R. Lazarus, "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, no. 1, p. 30, 2006.

[71] Open Health Natural Language Processing Consortium, "The MedKATp pipeline," www.ohnlp.sourceforge.net/MedKATp/, 2015, accessed: 2015-06-12.

[72] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[73] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[74] C. Manning and H. Schtze, *Foundations of statistical natural language processing*. MIT press, 1999.

[75] M. Marcus, M. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[76] i2b2, "NLP research data sets," www.i2b2.org/NLP/DataSets/Main.php, 2011, accessed: 2015-07-17.

[77] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley *et al.*, "Concept annotation in the CRAFT corpus," *BMC Bioinformatics*, vol. 13, no. 1, 2012.

[78] SURGbase, "SURGbaseEE operation report," www.surgbase.com.au/surgical-audit/features/surgical-logbooks.html, 2009, accessed: 2015-07-17.

[79] Y. Wang, S. Pakhomov, N. E. Burkart, J. O. Ryan, and G. B. Melton, "A study of actions in operative notes," in *Proceedings of the AMIA Symposium, American Medical Informatics Association*, vol. 2012, 2012, Conference Proceedings, p. 1431.

[80] Z. S. Harris, "The structure of science information," *Journal of Biomedical Informatics*, vol. 35, no. 4, pp. 215–221, 2002.

[81] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: a description based on the theories of Zellig Harris," *Journal of Biomedical Informatics*, vol. 35, no. 4, pp. 222–235, 2002.

[82] K. Verspoor, K. B. Cohen, and L. Hunter, "The textual characteristics of traditional and open access scientific journals are similar," *BMC Bioinformatics*, vol. 10, p. 183, 2009.

[83] Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[84] N. Bouayad-Agha, "The patient information leaflet (PIL) corpus. The Open University," 2012.

[85] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[86] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[87] P. Rayson and R. Garside, "Comparing corpora using frequency profiling," in *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics*, 2000, Conference Proceedings.

[88] K. Gorman, "Simpler sentence boundary detection," http://sonny.cslu.ohsu.edu/~gormanky/blog/simpler-sentence-boundary-detection/, 2014, accessed: 2015-09-15.

[89] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[90] A. R. Aronson, "The MetaMap mapping algorithm," National Institutes of Health (NIH), Tech. Rep., 2000.

[91] R. Batuwita and V. Palade, "Class imbalance learning methods for support vector machines," *Imbalanced learning: Foundations, algorithms, and applications*, pp. 83–99, 2013.