Submitted for examination of Doctor of Philosophy

A modular, open-source information extraction framework for identifying clinical concepts and processes of care in clinical narratives

Philip Gregory Gooch

December 5, 2012

Centre for Health Informatics

School of Informatics

City University London

Contents

List of Figures

List of Tables

Acknowledgements

My thanks to supervisors Professors Abdul Roudsari, John Chelsom and Andrew MacFarlane for their continuous support and valued advice over the past three years. Thanks are also due to Angus Roberts, Mark Greenwood, Ian Roberts, Diana Maynard, Valentin Tablan and Hamish Cunningham for technical help and advice on developing for the General Architecture for Text Engineering (GATE) platform; and to Rob Challen and Chris Wroe at the *British Medical Journal* for access to *Best Practice* and *Clinical Evidence* data.

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded in part by the grant number 2U54LM008748 from National Library of Medicine, and were originally prepared for the 2011 i2b2/VA Challenges in Natural Language Processing for Clinical Data Shared Task supported by the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the National Institutes of Health, National Library of Medicine under grant number R13LM010743-01.

Finally I would also like to thank my wife Sarah for her patience and encouragement over many long nights and weekends spent working on this project and associated publications. I could not have been completed this work without her support.

Declaration

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only copies made for study purposes, subject to normal conditions of acknowledgement.

I declare that this work and associated programming code, tables, figures and results are my own except where otherwise acknowledged and referenced. No part of this thesis has been submitted elsewhere for any other degree or qualification.

Abstract

In this thesis, a synthesis is presented of the knowledge models required by clinical information systems that provide decision support for longitudinal processes of care. Qualitative research techniques and thematic analysis are novelly applied to a systematic review of the literature on the challenges in implementing such systems, leading to the development of an original conceptual framework.

The thesis demonstrates how these process-oriented systems make use of a knowledge base derived from workflow models and clinical guidelines, and argues that one of the major barriers to implementation is the need to extract explicit and implicit information from diverse resources in order to construct the knowledge base. Moreover, concepts in both the knowledge base and in the electronic health record (EHR) must be mapped to a common ontological model. However, the majority of clinical guideline information remains in text form, and much of the useful clinical information residing in the EHR resides in the free text fields of progress notes and laboratory reports. In this thesis, it is shown how natural language processing and information extraction techniques provide a means to identify and formalise the knowledge components required by the knowledge base.

Original contributions are made in the development of lexico-syntactic patterns and the use of external domain knowledge resources to tackle a variety of information extraction tasks in the clinical domain, such as recognition of clinical concepts, events, temporal relations, term disambiguation and abbreviation expansion. Methods are developed for adapting existing tools and resources in the biomedical domain to the processing of clinical texts, and approaches to improving the scalability of these tools are proposed and evaluated. These tools and techniques are then combined in the creation of a novel approach to identifying processes of care in the clinical narrative.

It is demonstrated that resolution of coreferential and anaphoric relations as narratively and temporally ordered chains provides a means to extract linked narrative events and processes of care from clinical notes. Coreference performance in discharge summaries and progress notes is largely dependent on correct identification of protagonist chains (patient, clinician, family relation), pronominal resolution, and string matching that takes account of experiencer, temporal, spatial, and anatomical context; whereas for laboratory reports additional, external domain knowledge is required. The types of external knowledge and their effects on system performance are identified and evaluated.

Results are compared against existing systems for solving these tasks and are found to improve on them, or to approach the performance of recently reported, state-of-the-art systems. Software artefacts developed in this research have been made available as open-source components within the General Architecture for Text Engineering framework.

Glossary of abbreviations used in this thesis

General

API Application programming interface

CDS Clinical decision support

CDSS Clinical decision support system

CIG Computer-interpretable guideline

CPG Clinical practice guideline

CPOE Clinical physician order entry system

CSCW Computer-supported co-operative work

CUI Concept unique identifier

EHR Electronic health record (see also EPR)

EPR Electronic patient record

GATE General Architecture for Text Engineering

GL Guideline

HIS Health information system

HL7 Health Level 7

IE Information extraction

Glossary of abbreviations used in this thesis

NLM National Library of Medicine

NLP Natural language processing

OWL Web ontology language

PDF Portable document format

PN Petri Net

RDF Resource descriptor framework

RTF Rich text format

RIM Reference Implementation Model

SD System dynamics

SNOMED CT Systematized Nomenclature of Medicine – Clinical Terms

SOA Service oriented architecture

SSM Soft systems methodology

SWRL Semantic Web rule language

UMLS Unified Medical Language System

VMR Virtual medical record

Wf Workflow

WfMS Workflow management system

XML Extensible markup language

Part of speech tags for natural language processing

CC Conjunction

IN Preposition

JJ Adjective

NN Noun

NP Noun phrase

PP Prepositional phrase

PRP Pronoun

RB Adverb

VG Verb group

VP Verb phrase

1. Introduction

1.1. Background and motivation

Clinical decision support systems (CDSS) aim to provide patient-specific diagnostic and treatment recommendations to clinicians by matching information known about the patient to relevant medical knowledge residing in a repository or *knowledge base*[1]. CDSS have evolved over the years from standalone, single-domain[1] expert systems that, when provided with manually entered data, gave diagnostic and treatment suggestions, to more recent systems that monitor events in the electronic health record (EHR) and provide decision support in the form of reminders, alerts and advice on treatment and management derived from published clinical guidelines. Typically, such systems provide support for individual clinical decisions at a single point in time, but recently there have been calls for the development of *process-oriented* systems that provide support for longitudinal processes of care and decision-making that extend over time[2].

Peleg and Tu[3] describe two key knowledge management tasks in the development of a modern CDSS: 1) a requirements engineering task that involves identifying the processes of care, the goals, flow of information work activities required by the organisation and the roles and patterns of communication between care providers; and 2) a modelling task which involves representing clinical and organisational knowledge in a computer-interpretable formalism. In addition, for the CDSS to apply this modelled knowledge to data in the EHR, clinical terms and process knowledge concepts both in the knowledge base and the EHR need to be mapped to a common, controlled terminology[4][5].

In a recent review, Ahmadian et al.[6] found that a critical factor in the success of CDSS

[1] e.g. for differential diagnosis of abdominal pain or infectious diseases.

implementation was the availability of data mapped to a standard terminology. However, in the structured data entry environment of a typical EHR, data items required by CDSS are not always present, and it has been suggested that up to 50% of the useful clinical information resides in free text fields[7]. Demner-Fushman et al.[8] also noted that many opportunities for decision support can only be found in the free text of the patient record, such as described in history and examination or laboratory reports.

For some years, researchers have been calling for the availability of shared knowledge base components, and the tools for creating them. Clayton[1] argued that the lack of structured information, in the form of coded patient data and decision rules, was the main barrier to the uptake of CDSS (although, as we shall see in Chapters 2 and 3, there are many other factors to consider also). Greenes[9], when comparing the impact of search engines, such as Google, on the use of the World Wide Web to answer clinical questions, with the current state of CDSS, gave the following call to arms:

> 'Imagine the stimulus that a well-researched, evidence-based repository of knowledge, in a standardized, computable form, and tools for delivery of it in local settings in a patient- specific manner at times of need, would have on the ability to implement and demand for CDS capabilities'[9].
>
> 'A variety of clinical problems can be addressed simply by considering a knowledge base of rules. Error prevention/patient safety depends on rules that can be used to warn about potentially harmful actions, such as medical contraindications, and to alert providers to situations requiring action such as abnormal lab results. Best practice depends on rules that are used in actionable parts of guidelines that are translated into real-time treatment recommendations and alerts'[9].

and

> 'It may be desirable to maintain common repositories of computer-interpretable, unambiguous knowledge content (e.g. guidelines, or decision rules) for use across an enterprise'[10].

Greenes suggested that the lack of such a shareable, knowledge-based infrastructure has hampered the widespread usage of CDSS, as the majority of applications have been in a single host, application-specific environment[9]. Demner-Fushman et al.[8] argued that *natural language processing* (NLP) techniques have the potential to facilitate the knowledge base creation process, both in terms of extracting decision rules from clinical guidelines and extracting facts from the free text of clinical notes.

Sittig et al.[7] identified a number of 'grand challenges' in clinical decision support, which included the identification and classification of information in the free text of the EHR to drive clinical decision support. However, Fox et al.[11] raised concerns about the clinical safety of relying on information automatically extracted in this way, although Wagholikar et al.[12] have recently demonstrated an accurate cervical cancer screening CDSS that makes use of a free-text knowledge base as envisaged by Sittig et al., with the aim of guiding the physician with recommendations rather than completely automating the decision-making process.

The application of both statistical and lexical NLP techniques to extract terms and concepts in clinical narratives has a long history, starting with the work of Pratt and Pacak in 1969[13]. Interest in this area has grown rapidly over the past few years as a result of the increase in availability of anonymised clinical notes from US institutions for research purposes: for example, the i2b2 datasets[14], and in the UK there have been recent plans to make similar data available for research. Similarly, there has been an increase in clinical information available via the World Wide Web, in the form of guidelines available from the National Guideline Clearing House in the US, and the National Institute for Clinical Evidence (NICE) in the UK.

A number of previous research efforts have tended to focus on applications that extract terms and relationships from specific types of clinical texts (e.g. discharge summaries[15], radiology reports[16], and clinical guidelines [17]), but there have been few attempts to generalise a framework that can be used across each of these data sources (see Chapter 5). While there are a number of individual, open-source components for addressing each of these sub-domains, combining them together into a pipeline requires a large amount of

ad hoc 'glue' code[18]. Another of Sittig et al.'s[7] 'grand challenges' was the creation of shareable, 'plug and play' modules for CDSS. Therefore the challenge to develop shareable, interoperable components in a framework without requiring any 'glue' code is an additional motivating factor for this research.

1.2. Aims and objectives

The aim of this research is to build on existing methods for automated identification and classification of clinical concepts and processes of care from heterogenous textual resources (residing in clinical guidelines, protocols, free text clinical notes and research papers in a variety of text encodings and formats). By providing new methods and tools to do this, we aim to facilitate the knowledge formalisation process and the development of knowledge bases for process-oriented clinical decision support systems.

This research has the following objectives:

1. Identify the types of formalised knowledge required by health information systems (HIS) that provide process-oriented clinical decision support.

2. Identify the current challenges in implementing process-oriented HIS.

3. Develop a conceptual model for the development of process-oriented HIS, and identify where a framework for automated knowledge extraction and formalisation might sit within such a model;

4. Identify current research problems in the automated extraction of conceptual and process knowledge in the clinical domain.

5. Design, develop and evaluate an open-source[2], modular framework to solve a subset of these research problems and evaluate them along the following axes:

    a) performance in relation to accuracy in comparison with manually curated knowledge resources;

[2]Where 'open source' follows the Open Source Initiative definition of freely distributable software and source code: `http://opensource.org/osd.html`

b) performance in relation to previous approaches in terms of accuracy and speed.

Following Greenes and Sittig's suggestion for shareable, 'plug and play' modules, a related objective is that the developed framework should not be monolithic; it should consist of interoperable modules that can be swapped and configured for different clinical knowledge extraction tasks, with minimal configuration and without requiring programming expertise by the end user – a weakness of some existing open-source frameworks for processing clinical text, as discussed in Chapter 5.

1.2.1. Research hypotheses

The background, motivation, aims and objectives of this research lead to two hypotheses:

1. Complex clinical information extraction tasks can be assembled from linear pipelines of self-contained components, in which the output from component A may form the input to component B. However, component B should not require component A in order to complete its own subtask, only on the output of global components required by both A and B.

2. Such components can be created from external knowledge resources, and lexical and syntactic patterns derived from regular expressions operating over lexemes extracted from these knowledge resources.

1.2.2. Methodology and scope

A high-level overview of the research methodology employed in this research comprises: 1) systematic literature review and thematic analysis (Chapter 3); followed by 2) an iterative development cycle (Chapter 4) involving purposive sampling of external knowledge resources and representative documents to identify patterns (Chapters 5–8); 3) evaluation of system performance using publicly available 'gold standard' data sets in the clinical and biomedical domain; and 4) review of results against stated research objectives and results from previous research. Chapter 4 sets out the overall research and evaluation

methodology, and methods specific for individual framework tasks are detailed in their respective chapters that follow.

The aims and objectives allow us to limit the scope of this thesis to the clinical knowledge extraction, concept mapping and formalisation process. The output of individual components of the framework developed are evaluated against known 'gold standard' data sets; however the creation and validation of a formalised knowledge base for CDSS from the structured information extracted by this process is out of scope, but remains an area for future research (see Chapter 9).

1.3. Structure of this thesis

Chapter 2 gives an overview of approaches to modelling processes of care in the form of computer-interpretable guidelines and clinical workflows (which tend to reflect idealised processes and are external to the patient record), and care plans and care pathways (which aim to reflect patient-specific care processes and are part of the patient record). The aim of such models is to facilitate the provision of contextual clinical decision support at the point of care, i.e. tailored for the specific patient under consideration by the clinician during a consultation. The question of how these models are realised in practice is considered in Chapter 3, which reports on a systematic review of the literature on the challenges in implementing health information systems that provide process-oriented clinical decision support, and in which a conceptual model of the development process is proposed.

Chapter 4 selects one component of the model developed in Chapter 3 – the extraction and formalisation of clinical knowledge from text – as the focus for development. It provides a brief overview of NLP techniques in the clinical domain and details the research, development and evaluation methodology used in the remainder of the thesis. This method is applied in the development of a modular, open-source framework for extracting clinical concepts and process information from the textual data in clinical guidelines and patient notes, as described in Chapter 5. In that chapter, the core components of the framework are outlined and evaluated. Chapter 6 attempts to generalise some of the approaches of Chapter 5 into a more lightweight approach to clinical concept identification

and classification.

One clinical NLP problem not so far considered in the framework is the handling of abbreviations and acronyms. We can identify that the 'NLP' being discussed here refers to 'natural language processing', not 'neuro-linguistic programming' (a process known as *disambiguation*[19]), by either linking each mention of 'NLP', at the point at which it occurs in the document, back to its most recent definition in the text, or, in the absence of a such an expansion, select all its definitions from a dictionary and make use of contextual features in the document to select the best one. A method for implementing this is described and evaluated in Chapter 7, and is made use of in the ensemble pipeline in Chapter 8, which considers the problem of coreference resolution: the identification of textual descriptions that refer to the same real-world entity or event. In that chapter, it is argued that coreference resolution, in combination with identification of temporal expressions, events and their relations (developed in Chapter 5), is an important component in the identification of processes of care as linked chains of narrative events. In Chapter 8, a method for coreference resolution in the clinical narrative is developed, implemented as a pipeline process that makes use of all the framework components developed in the preceding chapters, and evaluated against a large corpus of clinical notes.

Finally, Chapter 9 discusses the results of the research in the wider context of identifying temporal processes of care, considers directions for future research, and draws conclusions.

The Appendices to this thesis contain additional material for Chapters 3 and 8. In addition, the CD that accompanies this thesis contains the data collection spreadsheet used for the systematic review (Chapter 3) and the software components developed in Chapters 5 to 8.

1.4. Contributions to knowledge

This thesis make contributions to knowledge in the following areas:

1. A qualitative meta-synthesis of research over the last 15 years into the modelling of clinical processes for decision support and a corresponding conceptual implementation framework (Chapter 3). This work was recently published in the *Journal of the*

American Medical Informatics Association[20].

2. Methods for adapting existing tools in the biomedical NLP domain to the task of processing clinical texts, and an evaluation of different approaches to improving the scalability of these tools when processing larger documents such as clinical guidelines (Chapter 5). Such approaches lead to linear scaling of processing time in relation to document size, rather than the exponential scaling currently encountered with some of these tools. This work was presented at the *Intelligent Data Analysis in Medicine and Pharmacology Workshop* at the *13th Conference on Artificial Intelligence in Medicine (AIME'11)*[21].

3. A method for identifying and expanding biomedical abbreviations that uses regular expressions dynamically generated from document content (Chapter 7). The method provides in-place annotation, expansion and coreference of abbreviations back to their initial or most recent definition in the text, in a single processing pass through each document. The method requires no training data; however, via runtime customisation of its input parameters it can be trained if required so that optimal parameter values can be calculated to tune the performance for different corpora[22]. In addition, a contribution to knowledge is made in terms of provision of corrected versions of two reference corpora used for evaluating the performance of biomedical abbreviation identification systems.

4. A method for semantic decomposition and rule-based recombination of ontology resources for simplifying the creation of biomedical and clinical concept recognisers (Chapter 6). In this thesis, the method is applied to identifying anatomical terms and disease concepts, and is evaluated against a corpus of manually annotated clinical notes. A contribution is also made in terms of a systematic quality assurance process for ontologies, which has allowed a number of errors in two reference standard biomedical ontologies – the Foundational Model of Anatomy and the Disease Ontology – to be identified, validated and corrected.

5. A method for resolution of coreference and anaphoric relations between terms oc-

curring in clinical notes, which makes use of the above methods and all the modules developed during this research (Chapter 8). The work was recently published in the *Journal of Biomedical Informatics*[23]. A contribution to knowledge is also made in terms of an analysis of the role of external domain knowledge resources in identifying these relations.

Part I.

Formalisation and implementation of care process models

2. Modelling processes of care

This chapter provides an overview of approaches to the modelling of healthcare processes at the clinical and organisational level. These types of representation aim to capture both declarative knowledge (definitions and statements of facts about concepts, their properties and relations) and procedural knowledge (sequences of tasks and rules that involve operations on declarative concepts). In the context of decision support, such representations aim to provide a formal definition of the clinical data items that form the input and output of the process, the flow of information, sequencing of tasks, and the roles of the participants that will perform the tasks involved in the care process. In addition to the ability to support individual clinical decision making at the point of care, the goal of such formalisms is to support the treatment and management of care processes that extend over time.

Terminology often used to describe processes of care in the context of clinical decision support include clinical guidelines, protocols, care pathways, care plans, and clinical workflow. This chapter shall define, compare and contrast these terms (Section 2.1) and will consider the different types of formalisms that have been used to model the healthcare processes to which they refer (Sections 2.2 and 2.3). Moreover, they all sit within the overarching concept of *workflow*, so we begin with a definition of this term.

2.1. Definitions

2.1.1. Workflow

Workflow involves the sequencing of tasks and flow of information in an organisational process (clinical or otherwise) and is defined by the Workflow Management Coalition

(WfMC) as:

> The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of *procedural rules*.[1]

In addition:

> The automation of a business process is defined within a *process definition*, which identifies the various process activities, procedural rules and associated control data used to manage the workflow during process enactment.[1]

and

> A *Workflow Management System* (WfMS) consists of software components to store and interpret process definitions, create and manage *workflow instances* as they are executed, and control their interaction with workflow participants and applications.[1]

(my emphasis added)

In a workflow system, the definition and execution of the appropriate control and data flow, the assignment of people to tasks and the invocation of the application logic blocks (workflow *execution*) are separate from the application logic (programming code) itself. Changes to the process can therefore be made without impacting the application logic[2].

2.1.2. Clinical workflow

Niazkhani et al.[3], in developing a conceptual model based on the principles of workflow and computer-supported co-operative work (CSCW), defined clinical workflow as

> the flow of care-related tasks as seen in the management of a patient trajectory: the allocation of multiple tasks of a provider or of co-working providers in the processes of care and the way they collaborate[3]

Niazkhani et al. categorise clinical workflow into four inter-related and inter-dependent elements:

1. *structuring of clinical tasks*: integration of domain and healthcare knowledge (what, when, where, who);

2. *co-ordination of work*: scheduling, synchronisation, roles, resource allocation, temporal constraints

3. *information flow*: integrating expertise, guidelines and protocols with knowledge from the medical record

4. *monitoring*: making dynamic changes to clinical tasks in the light of new information

2.1.3. Clinical decision support systems

Clinical decision support systems (CDSS) aim to provide diagnostic and treatment recommendations and advice at the point of care, i.e. information tailored for the specific patient under consideration by the clinician at a given moment[4], in order to improve practitioner performance, the quality of care, and better patient outcomes, as a result of more informed, evidence-based decision making[5]. Such systems can be classified as *active* and *passive*[5][6]. Active systems provide automated advice in the form of alerts, commentary and recommendations in response to events occurring within the application while the user works. For example, following the entry of a medication order into a computerised provider order entry (CPOE) system, the CDSS may automatically check for potential contraindications or unwanted drug interactions[5]. Alternatively, an active CDSS may automatically provide treatment recommendations following the availability of new data in the electronic health record (EHR) (e.g. demographic information, family history, vital signs, laboratory results). In a passive CDSS, however, the user is required to manually invoke or consult the system first before receiving decision support. One common method of implementing a passive CDSS is via *infobuttons*[7], where the clinician invokes a contextual information button that sends a message payload, containing parameterised user and patient data (e.g. details of the clinical task being performed; patient gender, date of birth, diagnosis) as a search query to an online knowledge resource, which returns relevant results.

A systematic review of the features of clinical decision support systems critical for improving clinical practice[8] found that active rather than passive CDSS were more likely to improve practice (as measured by patient outcomes or measures such as guideline adherence or reduction in prescribing errors). In addition, they found that the provision of actionable recommendations via a computer, at the point of care, and integrated with clinical workflow were key success factors in a CDSS.

2.1.4. Clinical guidelines

Clinical guidelines (also known as clinical practice guidelines and clinical protocols) have been defined as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances"[9]. Such statements contain recommendations for best practice based on systematic reviews of clinical evidence, consensus statements and expert opinion. The goal of guidelines is to reduce variation in medical care by promoting the most effective treatments, and to provide a means of quality control in clinical practice (for example, auditing treatment provide against that suggested by guidelines).

The use of and adherence to guidelines by clinicians is often claimed to be poor[45]. A number of reasons have been suggested for this:

1. Information overload: the sheer number of guidelines available, and the amount of time and effort required to absorb the information contained within them[46][47]. In the UK alone, best practice guidelines for a large number of conditions are available from a variety of sources; for example, the Scottish Intercollegiate Guidelines Network (SIGN); NHS Clinical Knowledge Summaries; the National Institute for Clinical Excellence (NICE); the Royal College of Physicians, among others.

2. The fact that most guidelines are available only as free text, making it difficult for the clinician to find the most appropriate information[46].

3. Guidelines may either be too general, ambiguous, incomplete or they may be too specific, making them hard to fit to local practice[47].

4. Guideline recommendations may not fit the actual flow of patient encounters or clinical workflow in an organisation[48].

5. There may be a mismatch between guideline recommendations and the clinician's mental model of what is appropriate care for a specific patient[49].

6. Guidelines do not necessarily provide the knowledge and information required in order to implement them in practice[50].

Goud et al.[51], following the work of Cabana et al.[52], summarised these barriers to guideline use in practice as being *internal* and *external*. Internal barriers include the clinician's attitude towards and knowledge of the guideline content. External barriers include the guideline content itself (e.g. complexity, consistency, ambiguity); patient factors (e.g. consent to treatment); and organisational factors (e.g. lack of time and resources). Carefully planned organisational change is required to overcome these barriers for successful guideline implementation[51].

The 'computerisation' of clinical guidelines (e.g. via the Web or on CD-ROM) has been proposed as a method for removing some of the barriers to adoption and use of guidelines[53]. However, simply providing clinical guidelines in an electronic format is not sufficient to lead to improved decision making or guideline adherence[54][55][56]. Instead, guidelines need to be seamlessly integrated into the existing health information system (HIS) or electronic patient record (EPR), integrated with day-to-day clinical workflow, tailored to the individual patient and be available at the point of care[55][57][58].

Systematic reviews have shown that when clinicians make use of clinical decision support systems that include a knowledge-base derived from clinical practice guidelines, their adherence to guidelines is improved[59]. However, the effect on patient outcomes of such systems is mixed. A randomized-controlled trial of guideline-based care suggestions presented to physicians when writing orders for the treatment of patients with chronic respiratory diseases found no effect on patient outcomes[60], echoed also general finding of a recent systematic review.[61]

Shiffman et al.'s[62] systematic review of the functionality and effectiveness of guideline-

based CDSS found that such systems can improve guideline adherence and quality of documentation. However, the wide variety of system design, level of description, guideline implementation strategy and clinical setting made it difficult to determine which factors were important in influencing the system's success or failure. A more recent systematic review suggested that overall processes of care (e.g. information sharing, retrieval and automated provision of advice) are improved by guideline-based CDSS[63], although nearly 40% of the studies selected were reported by the system developers, and what was meant by process of care was different for each study, being a post-hoc binary intervention variable based on the conclusion of each study. Neither review distinguished between systems that simply present guideline-based recommendations on a computer – i.e. for individual clinical decisions – from systems that model and support longitudinal, longer-term clinical processes. Section 2.3.2 gives an overview of some of these models.

2.1.5. Care plans and care pathways

Care plans, or treatment plans, are "plans of future activities, specific to a patient's problem(s), treatment and goals, which are signed and time-stamped"[10]. Care plans may be discipline-specific (e.g. a nursing care plan), or multidisciplinary, but in both cases they are based on a full assessment of the patient's needs, and form a goal-directed treatment plan on how those needs are to be addressed. These plans form part of the patient record in the EHR, but may be referred to, in whole or in part ,in the free text of the patient's progress notes.

In contrast there is no single, agreed definition of a 'care pathway', and despite many years of use, the concept is somewhat unclear[11], often being used interchangeably with care plans, clinical guidelines and protocols[10]. The critical path method (see Section 2.2.2) used for project management seems first to have been applied to the management of clinical processes in the US by Zander[12] to improve the quality and efficiency of patient care, and the term 'critical pathways' was coined in relation to care processes. This evolved into the use of the term 'care maps' and 'clinical pathways' as outcome measures for planned processes of care were introduced[13]. In the UK, the term 'integrated

care pathway' (ICP) tends to be used. A widely cited definition of integrated care pathways (referred to here as 'care pathways' hereafter) by Campbell et al. [14] neatly links the concepts of care plans, clinical guidelines, protocols and workflow with the idea of a process-oriented patient record:

> ... structured *multidisciplinary care plans* which detail essential steps in the care of patients with a *specific clinical problem* ... They offer a structured means of developing and implementing *local protocols* of care based on *evidence based clinical guidelines*. They also provide a means of identifying the reasons why clinical care falls short of adopted standards. [They] describe the *tasks* to be carried out together with the *timing* and *sequence* of these tasks and the discipline involved in completing the task. They consist of a *single multidisciplinary record* which is part of the patient's clinical record.[14]

(my emphasis)

Care pathways are considered to be a complex intervention[11]: a recent systematic review[15] noted the poor quality of reporting of the care pathway implementation process which prevented analysis of which factors were critical to their success or failure, despite finding that the use of care pathways is associated with improved process of care (as measured by reduced in-hospital complications and improved documentation) but without increasing length of stay or hospital costs.[1] Perhaps one of the reasons for lack of reporting of the implementation process is that there are differing views on what a care pathway should contain and how it should be developed. Further confusion arises when the term is used to describe and model higher-level care commissioning processes, patient flows[16], or paths to care and routes of referral[17][18]. This is discussed further in [19] and in Section 2.3.5 below.

[1] Such findings might be considered to be unsurprising given that care pathways tend to focus on patients with a single, well-defined condition[11], but that is a discussion for another report.

2.2. Modelling organisational workflow processes

In order to automate a process, the components of that process, and the dependencies between them, must be formally defined and represented in a way that allows operations to be performed on them, and on the system as a whole. This section provides an overview of some of the ways in which workflow processes are modelled.

2.2.1. Discrete event models for workflow

In the workflow view of the world, processes consist of discrete events that occur in some scheduled order – which may be defined at design-time or may change as the process unfolds, depending on some predefined constraints being satisfied. A number of notations and formalisms have been developed for the purpose of modelling these events and their relationships.

Notations of various levels of formality include:

- *Business Process Modelling Notation* (BPMN) — an industry standard notation from the Object Modelling Group (OMG) and WfMC for visualising workflow process definitions
- *XPDL*[20] — an XML serialisation format for BPMN that also defines executable properties of the workflow
- *Event-driven Process Chains* (EPC)[21] — a graphical business process description language consisting of functions, events and logical connectors developed for use in the ARIS business process modelling framework.
- *YAWL*[22] — a workflow language based on a rigorous analysis of existing workflow management systems and workflow languages, using the semantics of Petri Nets (see below) as a starting point.

General, mathematical formalisms that underpin some of the above workflow notations include:

- *Petri Nets* (PNs)[23] – a formalism comprising finite collections ('bags' – sets in which duplicate elements are permitted) of places, transitions, and input and output functions that define mappings from transitions to places and vice versa. PNs can be visualised as a bipartite directed graph, i.e. with nodes consisting of places and transitions, where every place is connected to at least one transition and vice versa. PNs were originally designed to model concurrent interacting processes, and have been proposed as a suitable formalism for workflow modelling by van der Aalst[24]. A change of state is represented by the movement of tokens between an input place and an output place via an enabled transition.

- *Finite State Automata* (FSA) – a formalism comprising a finite set of states, and a transition function that maps the transition from one state to another based on an input alphabet of symbols. FSA can be visualised as a state transition table or as a directed graph. Their use as a formal model in workflow systems has been developed by Wombacher et al.[25], in which the input alphabet represents the possible events or messages that can be handled by the workflow.

- *Temporal logic*[26] – an extension of predicate logic with 'necessity' and 'possibility' temporal modifiers. In temporal logic based workflows, task ordering and control flow is not predefined, but are scheduled for execution when they satisfy some global dependencies. However the computational costs of verifying such workflows has been criticised as being too high[26]. Temporal logic features in computer-interpretable guideline formalisms such as Asbru, GLARE[27] and PROforma[28] (see Section 2.3.2), and guideline temporal constraints can also be modelled with an extended PN model known as 'coloured' Petri Nets[29]. Clinical guidelines encoded in the Asbru and GLARE formalisms have been verified using temporal logic model checkers[30][31], in which logical dependencies are decomposed into finite state automata.

- *Transaction logic*[26] – an extension of predicate logic comprising declarative and procedural operators for specifying state changes in logic programs and when mod-

elling database operations. Transaction logic can be used to model workflow as a series of database transactions, and has been used to simulate and provide proofs for workflow processes[26].

The Workflow Patterns Initiative[32] aims to formally describe and systematise the types of process control flow constructs that workflow languages and systems should support. These constructs are described by *patterns*: the abstraction of recurring forms that appear in a range of contexts. Forty-three control flow patterns have been identified, grouped as follows:

- Basic control flow (e.g. sequence, parallel split, exclusive choice)
- Advanced branching and synchronisation (e.g. multi-choice, multi-merge)
- Iteration patterns (e.g. arbitrary cycles, structured loops)
- State-based patterns (e.g. deferred choice, milestones)
- Cancellation patterns (cancel task)
- Termination patterns (implicit and explicit workflow termination)
- Trigger patterns (task external activation)

Petri Nets are considered to be particularly suitable for modelling workflow processes for the following reasons[33]:

- They combine a formal semantics with a graphical representation. This provides an unambiguous, tool-independent representation with reasoning and mathematical proof properties. PNs can model simple workflow primitives (AND/OR joins and splits, iteration) as well as the more complex workflow patterns described above.
- They allow both state-based and event-based execution. This allows a distinction to be drawn between the enabling of a task and the actual execution of a task: important for modelling temporal constraints, delays, and manual task execution.

- Numerous analysis techniques are available, such as detection of unreachable states, of deadlock, and confirmation that the workflow will always terminate eventually (i.e. that the workflow is 'sound')[33].

Figure 2.1 shows one of the basic control flow patterns, the parallel split, modelled as a Petri Net. For examples of other workflow patterns, including animations, see `http://www.workflowpatterns.com/patterns/control/`.

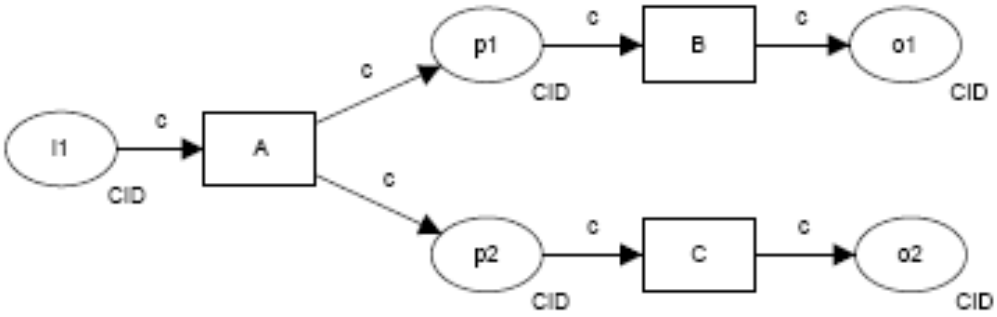


Figure 2.1.: Flow of control splits into two parallel threads in state p1–p2, allowing both tasks B and C to be activated
i1 is the input place (signalling the entry point of the workflow fragment), o1 and o2 are output places (signalling the termination of the workflow fragment), c represents some transition condition, CID represent the case identifier of the workflow instance.
Source: http://www.workflowpatterns.com/patterns/control/

2.2.2. Systems models of organisational processes

In many real-world systems, boundaries between tasks, roles and organisational groups may overlap. There may be a number of workflows or activity paths occurring simultaneously. There may be separate information flows, not connected to discrete workflow tasks and processes. Processes may also be continuous and dynamic; we might be interested in overall rates of flow of quantifiable things – for example, the flow of patients. A number of discrete and continuous models for representing dynamic systems have been developed.

PERT/CPM

PERT (program evaluation and review technique) and CPM (critical path method) are closely related, but independently developed, network techniques for planning and coor-

dinating project activities, developing and monitor a project schedule[34]. The features of both techniques have been combined into the PERT/CPM method for project management.

With PERT/CPM, an activity-on-node project network is created as a directed, acyclic graph consisting of nodes, representing project activity completion milestones (*events*), and arcs, representing the *activities* themselves that lead toward milestones, with the direction of the arc indicating the precedence relationship between milestones (Figure 2.2).

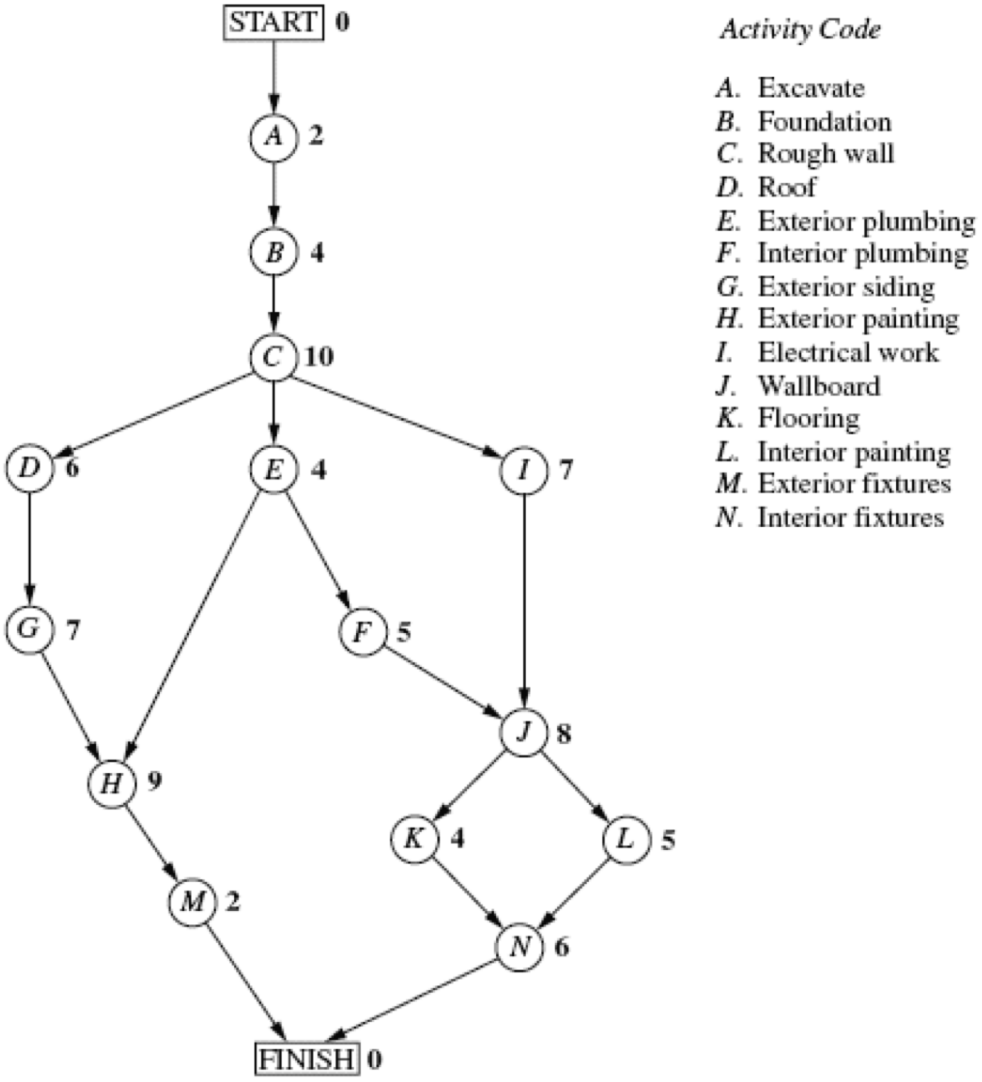


Figure 2.2.: Activity-on-node network
Source: Hillier and Liebermann[34]

The estimated duration of each activity is assigned to each node. The start and finish times of each activity, if no delays occur in predecessor activities, can be represented by the earliest start time (EST) and earliest finish time (EFT) of the activity. EST and EFT times are found by making a forward pass through the network. Similarly, the latest start (LST) and latest finish times (LFT) are the latest possible activity start and finish times that do not delay the overall project completion, and are found by making a backward pass through the network. In the example in Figure 2.2, completion of all activities requires ABCEFJLN = 44 days (as ABCDGHM=40, ABCIJLN=42, ABCEFJKN=43, ABCIJKN=41). The EST for activity H–M is via ABCDGH=2+4+10+6+7=29 days, its LST is 44-2-9=33 days.

Gantt charts

Gantt charts[35] provide a way of visualising an instantiated PERT/CPM network (from a given start day and date) after earliest and latest start and finish times for each activity have been calculated. The chart allows a timeline of execution of parallel and sequential tasks to be visualised, as well as the dependencies between them (Figure 2.3).

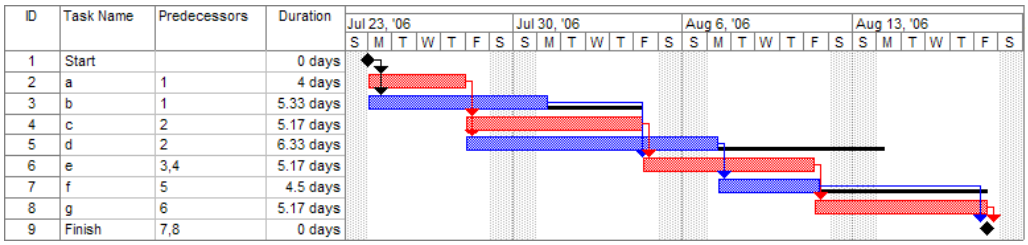


Figure 2.3.: Gantt chart for visualising task dependencies
Source: http://en.wikipedia.org/wiki/File:Pert_example_gantt_chart.gif

The activity-on-node representation of PERT/CPM and Gantt charts formed the basis of Chu's[36] care pathway model, and the EST/EFT temporal constraints of these models are formalised in the Asbru computer-interpretable guideline formalism[37] (see Section 2.3.5 below).

System dynamics

A system can be defined as a collection of parts organised for a purpose[38]. System dynamics (SD) is a branch of management science that deals with the dynamics and controllability of managed systems (i.e. those that are influenced by the actions of people). Managed systems make use of policies (inputs, controls) to control the system behaviour as time passes and circumstances change, making use of feedback loops that affect rates of change in the system variables[54]. The temporal aspect of system dynamics distinguishes it from other approaches such as decision theory, and the use of continuous variables that model rates of change and rates of flow distinguishes it from discrete event modelling formalisms such Petri Nets.

A key part of system dynamics is the use of the *influence diagram* or causal loop diagram. An influence diagram shows flows into and out of parts of the system, and the variables that affect this flow. Solid lines show physical flows — the consequences of actions. A '+' means that as the variable at the tail of the arrow increases, the variable at the head changes in the same direction; a '–' means that as the variable at the tail of the arrow increases, the variable at the head changes in the opposite direction (Figure 2.4). A simulation model is then constructed using a *stock-flow diagram* (Figure 2.5). If the influence diagram has been drawn correctly, the stock-flow model can be written from it directly. Software is available to assist in this step.

Brailsford[16] suggests some reasons why system dynamics is useful for modelling healthcare systems:

- SD focuses on system structure and drivers for change;
- SD models are high level models that use aggregate functions, and do not require large amounts of individual data items;
- SD can represent large, complex systems whose boundaries overlap with other organisations.

Examples of possible uses of SD for modelling organisational and managerial processes in healthcare include:

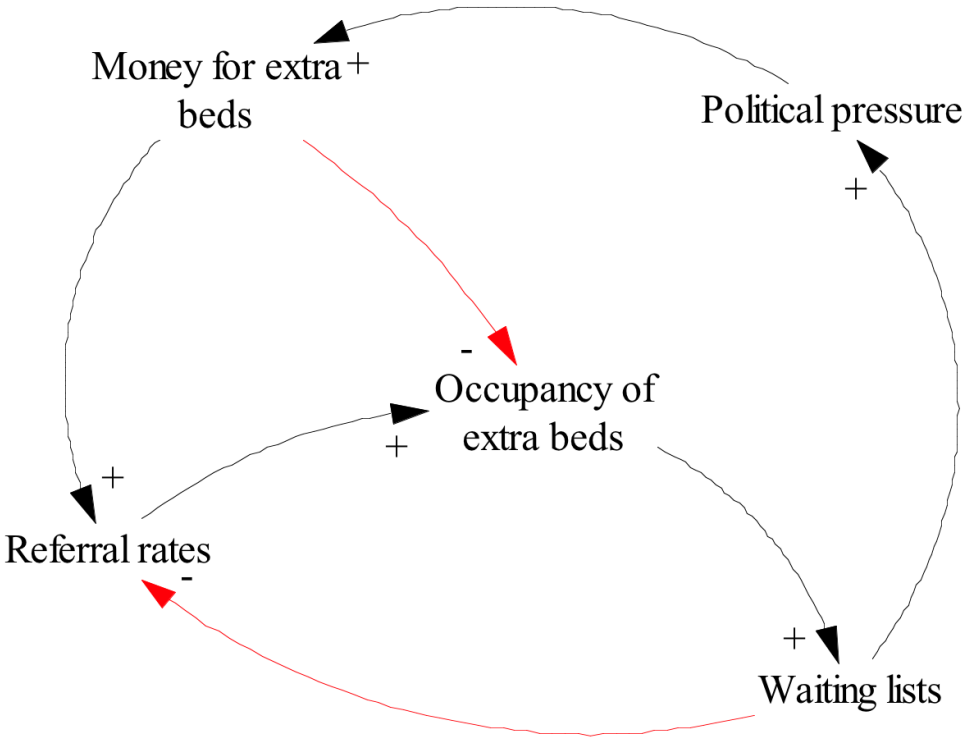


Figure 2.4.: Example influence diagram for hospital admissions
Source: Brailsford[16]

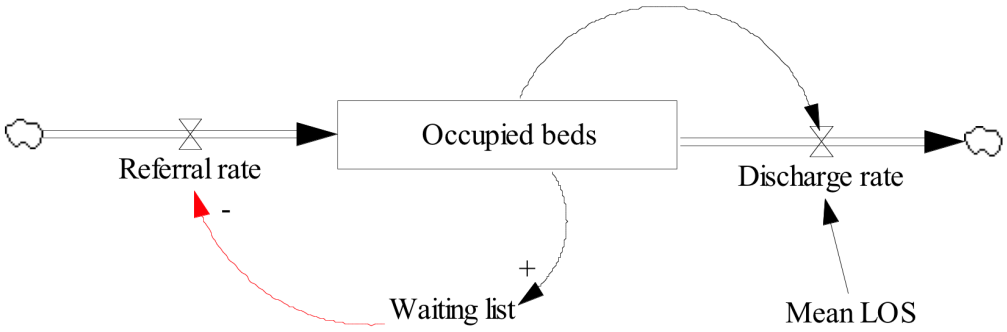


Figure 2.5.: Example stock-flow diagram for hospital admissions
Source: Brailsford[16]

- determining optimal patient flows into referral/treatment, given varying waiting list times and discharge delays;
- determining the optimal number of acute beds required in order for the organisation to meet a given protocol target;

- determining the shortest initial assessment waiting time required in order to provide a given referral to treatment time exploring the impact of changing antidepressant prescribing rate on patient recovery.

Soft systems methodology

Soft systems methodology (SSM)[39] is an approach to systems thinking that focuses on the human and social factors involved in organisational change. In contrast to the reductionist approaches that might be involved in formal modelling of deterministic workflow systems, SSM aims to develop a consensus view of the whole system that is greater than the sum of its parts. It is designed to model complex organisational systems with a plurality of viewpoints — features that are typical of a healthcare organisation.

SSM involves[40]:

1. Building a 'rich picture' of the organisation, in collaboration with the various stakeholders. This will be a conceptual map that may share features of the influence diagram used in system dynamics.

2. From the rich picture, one or more root definitions are developed that describe the system under study.

3. For each root definition, perform a *CATWOE* analysis of the stakeholders and processes involved:

    - **C**ustomer: the beneficiaries of the system
    - **A**ctor: who performs the activities in the system
    - **T**ransformation: the process - what input is transformed into what output?
    - **W**orldview: the vision of what the process should achieve
    - **O**wner: the owner of the process, who has the decision to change or abolish the system
    - **E**nvironment: the environment in which process takes place or that is assumed by the system

4. Develop a conceptual model for the Transformation (T) described in CATWOE in terms of the set of activities required to execute T.

5. Compare the conceptual models with the rich picture developed in Step 2 and develop an agenda of possible changes to the system.

6. Select desirable and culturally feasible changes that may be implemented.

7. Implement the changes, which may involve the use of hard/reductionist systems methodologies.[40]

The following section considers how these various formalisms and approaches, including combinations and variations thereof, have been used to model clinical processes of care.

2.3. Modelling clinical processes

In this section we discuss approaches to modelling planned processes of care using approaches that combine the procedural knowledge expressed in a workflow model with the declarative knowledge expressed in a clinical taxonomy or, more formally, an ontology. An ontology provides a conceptual representation of knowledge within a given domain. It comprises a controlled vocabulary of concepts, their properties, relationships and restrictions. Such relationships include type (class membership), intersection, exclusive disjunction, meronymy (whole–part) and synonymy. These relationships can be used to describe and carry out inferential reasoning about the domain.

Ontologies also provide a means of mapping concepts from one domain to another, to facilitate data sharing and interoperability. In the biomedical and clinical domains, there have been efforts to create various interoperable, open-source reference ontologies, such as the Foundational Model of Anatomy (FMA)[41], the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[2], and Logical Observation Identifiers Names and Codes (LOINC)[3]. The National Library of Medicine's Unified Medical Language System (UMLS)[4] attempts to integrate these separate resources into a *Metathesaurus*, where

[2] http://www.ihtsdo.org/snomed-ct/
[3] http://loinc.org/
[4] http://www.nlm.nih.gov/research/umls/

terms common to a number of individual ontologies are mapped to a single, canonical form. These ontologies are curated at a number of online libraries, such as the BioPortal at the US National Center for Biomedical Ontology, and the Open Biomedical Ontology Foundry.

The World Wide Web Consortium (W3C) has developed the Web Ontology Language (OWL) formalism to provide a means for developing ontologies that can be shared over the World Wide Web, and allows the creation of larger ontologies that can reference or import component ontologies via the use of namespaces. OWL ontologies are a core component of the Semantic Web[42], which aims to create a web of interoperable, reusable data from the web of unstructured documents. As we will discover in Chapter 3, the use of OWL and other semantic web formalisms is becoming increasingly important in the modelling and execution of clinical process workflows.

Given that we are concerned with care processes that extend over time, this section begins by looking at how concepts of time and temporal relationships between concepts have been modelled by researchers.

2.3.1. Formalising temporal knowledge for clinical processes

In Section 2.2.2 we considered how temporal constraints between tasks such as earliest start and finish times could be calculated from estimated task durations and traversing the directed graph of activities and events in a forward and backward direction. However, representing temporal clinical knowledge in terms of such a directed, acyclic graph is limited as it does not represent repetitions or periodicity that typify many real-world clinical processes – for example, a treatment regime consisting of therapy repeated 6 times where each cycle of therapy last 5 days, with a defined delay between each cycle, and where on each therapy day medication is administered every 12 hours[43].

TimeML[44] is an emerging standard for capturing and reasoning with temporal expressions occurring in narrative text. As a markup language (i.e. containing elements that wrap textual phrases, with information about the concept metadata stored as attributes of the element), it allows events described in documents to be annotated with respect to

their anchors in time, relative ordering and context (e.g. '*in 2 weeks*'), and persistence. TimeML has separate representation primitives for the identification of events (`<EVENT>`), temporal expressions (`<TIMEX3>`), and the relationships between events or between an event and a temporal expression (`<TLINK>`).

The TimeML specification is fairly complex and detailed, so only a few examples applied to the clinical domain will be given here (see Chapter 5 for an application of TimeML to clinical discharge summaries). From a clinical perspective, a TimeML `<EVENT>` would be a clinical concept (e.g. '*type 1 diabetes mellitus*', '*renal function*'), a verb or verb group representing a process (e.g. '*should be referred*', '*will be discharged*'), or a concept or process modifier (e.g. '*decreasing*', '*severe*'). A `<TIMEX3>` annotation captures specific, relative or approximate dates, durations and frequencies. For example, the approximate duration

> for at least 3 days

would be specified in TimeML as

```
<TIMEX3 tid="t0" type="DURATION" value="P3D" mod="EQUAL_OR_MORE">3 days</TIMEX3>
```

Optional `beginPoint` and `endPoint` attributes can be used to anchor the duration to specific dates that are not available at workflow design-time, but only at run-time. For example, if a process is begun at 10pm on 29 July 2012, represented as

```
<TIMEX3 tid="t1" type="DATE" value="2012-07-29T22:00">
```

then the above expression would become:

```
<TIMEX3 tid="t0" type="DURATION" value="P3D"
mod="EQUAL_OR_MORE" beginPoint="t1">3 days</TIMEX3>
```

Cycles and repetitions are dealt with by the `SET` type. For example

> Once a day for 3 days each week for 4 weeks

would be specified in TimeML as

```
<TIMEX3 tid="t2" type="SET" value="P1D">Once a day</TIMEX3>
<TIMEX3 tid="t3" type="SET" value="P1W"
quant="EACH" freq="3D">3 days each week</TIMEX3>
<TIMEX3 tid="t4" type="DURATION" value="P4W">4 weeks</TIMEX3>
```

While the above – and more complex – expressions of repetition and periodicity can be represented in TimeML, one potential problem with the specification is that its primitives are anchored to natural language. For example, the `SET` type can specify a repetition, but the specification only allows `beginPoint` and `endPoint` attributes to be used for `DURATION` types. Anselma et al.[43] sought to more formally specify the type of constructs required by care processes, and that may require more complete specification than TimeML-annotated clinical texts. They defined a `Repetition` primitive with more fully specified attributes than the equivalent TimeML `SET` type, such as repetition conditions `while()` and `onlyIf()`, `fromStart()` and `inBetween()` constraints that allow minimum and maximum delay times to be specified from the start of the first repetition, and between subsequent repetitions.

Anselma et al.'s formalism aimed to provide a method for both specifying and checking the consistency of temporal constraints defined in clinical guidelines. More generally, *computer-interpretable clinical guideline* (CIG) models aim to specify guideline processes, concepts and their dependencies, to allow both the clinical and temporal aspects of the care process to be captured and reasoned with computationally. The following sections provide an overview of these models.

2.3.2. Modelling clinical guidelines

In an effort to standardise the design and development of guideline-based CDSS, several formalisms for encoding content into a CIG format have been proposed. A number of comparative analyses of the most developed formalisms have been published[64][65][66]. Tu et al.[67] identified a typology of guideline modelling formalisms:

1. end-to-end flowcharts for algorithmic problem-solving processes consisting of 'IF ... THEN ... ELSE' statements for simple yes/no decision trees (e.g. see [68] and [69]

for example implementations);

2. decision maps – sets of patient scenarios forming a transition network of decision points, without specific start or end points;

3. partially ordered activities in care plans (see Section 2.1.5 and 2.3.4) that aim to meet defined goals (e.g. the Asbru formalism[70]); and

4. workflows (see Section 2.1.1 and 2.2.1) that take an organisational and role-based view of care processes (e.g. the GUIDE[71] and the ADEPT[72] systems).

More recent modelling formalisms take a hybrid approach that integrates these processes, such as GLIF3[73] and PROforma[28].

Tu et al.[74] proposed that guidelines either be modelled as decision maps for individual clinical decisions or as workflow processes that extend over time:

> 'We propose that recommendations in a clinical guideline can be structured either as collections of decisions that are to be applied in specific situations or as processes that specify activities that take place over time. We formalize them as "recommendation sets" consisting of either Activity Graphs that represent guideline-directed processes or Decision Maps that represent temporal recommendations or recommendations involving decisions made at one time point. We model guideline processes as specializations of workflow processes.'[74]

Identifying and extracting these 'recommendation sets' and their selection criteria from the text of the guideline may be possible with natural language processing techniques[75] (see Chapters 4 and 5), potentially providing a mechanism to reduce the information overload and information retrieval barriers that free-text guidelines present, or as an intermediate step towards guideline formalisation[76].

Broadly, the published guideline modelling formalisms fall into four categories:

1. Rule-based: e.g. Arden Syntax[77]

2. Document-based: e.g. GEM[78]

3. Decision-logic expression languages: e.g. GELLO[79]

4. Task-network models: GLIF[73], PROForma[28], SAGE[80], Asbru[70], GUIDE[71]

Standards exist for the first three categories: Arden Syntax, GEM and GELLO have been adopted as standards by HL7. Arden and GELLO are American National Standards Institute (ANSI) standards; GEM is an American Society for Testing and Materials (ASTM) standard. None of these, however, are sufficient for representing dynamic, complex care processes:

1. Arden Syntax is able to represent individual, independent clinical rules as Medical Logic Modules (MLM)[77] but lacks a defined semantics for representing clinical guidelines that are more complex than individual IF...THEN rules[28]. However MLM modules can be chained (one module invoking another if certain conditions are met) or invoked as sub-routines, although this requires a MLM composition logic to be defined, which simply defers the problem[81].

2. GEM provides a formalism for structuring complete guideline documents[50], but does not, on its own, provide a mechanism for execution or rule inferencing.

3. GELLO[[79] is query language that uses an object-oriented data model based around HL7 Reference Implementation Model (RIM) concepts to form a 'virtual medical record' (VMR). Its purpose is to provide a platform- and vendor-independent language for extracting, manipulating, and reasoning about data from medical records. Essentially, it provides an interface between the guideline (encoded in some process-oriented formalism) and the medical record.

The fourth type of formalism, which is yet to be standardised, specifies a process-flow-like model in which guidelines are composed of a network of tasks that unfold over time[82] and aim to support the type of process-oriented decision support described earlier. Although their syntax, semantics, and individual focus differ, these 'task network models' share a common set of features[64][65][66][82]:

- Decomposition of guideline concepts into a finite set of task-based *primitives* representing individual guideline steps, such as clinical actions, decisions, enquiries, and logical branch and synchronisation steps.
- The use of an *ontology* to classify the hierarchy of primitives, their attributes, and control relations that define their sequencing.
- The ability to organise the primitives into a clinical plan of nestable components, and the creation of sub-plans and sub-guidelines.
- The assignment of state to guideline steps, for example, in-progress, suspended, completed, cancelled, and conditions on the transitions between these states.
- The ability to define control flow: sequential, parallel and iterative task execution.
- The ability to define guideline entry points, and guideline step pre-conditions and post-conditions.
- The ability to specify temporal constraints on guideline steps.
- A graphical editor for creation of instances of encoded guidelines.
- An *execution engine* for enacting the encoded guideline instance.

One other feature these formalisms share is the ability to express a number of the standard workflow patterns summarised in Section 2.2.1. A study by Mulyar et al.[82] considered the expressive power of some the major CIG formalisms (PROforma, Asbru, EON, GLIF) as measured by the number of control-flow workflow patterns that they supported. They found that although these formalisms supported only around 50% of the workflow patterns, they offered unique features not found in current WfMS. These included the ability to model complex decisions via argumentation rules (rule out activity x, rule-in activity y) or expression languages, and the ability to specify multiple entry and exit points. However, they suggested that the CIG community might be advised to use more general information-driven workflow formalisms that allow more flexible execution, rather than construct specific languages for modelling clinical processes[82].

2.3.3. Modelling clinical workflow

Dadam et al.[72] first proposed the use of a WfMS to support and manage clinical workflow. While they accepted the need not to impede existing work processes and to allow flexibility in the selection and execution of clinical activities, they acknowledged the difficulty in supporting dynamic changes in a system where there may be complex interdependencies between tasks, given such sequences of tasks have been explicitly modelled. In practice, it was only possible dynamically insert a new task that was not dependent on others.

In a widely cited and seminal paper, Quaglini et al.[71] described a methodology and system architecture for 'careflow' (Figure 2.6): an integration of a CIG model for clinical tasks with a commercial workflow engine for managing organisational processes, and applied this to the management of acute stroke. Their system allowed for dynamic modifications (*exceptions*) to the predefined process by allowing tasks to be omitted or substituted for others. Schadow et al.[83] suggested that WfMS would be best used for well-defined and standardised clinical processes, such as immunization or clinical trials. Though neither Schadow or Quaglini used the term, there is much similarity between their approaches and the concept of a computerised care pathway, and we discuss this further in Section 2.3.5.

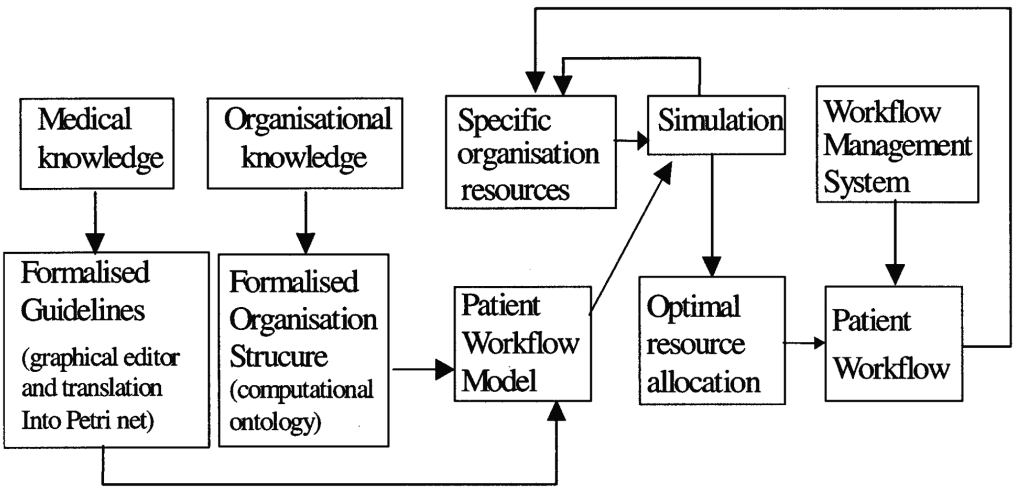


Figure 2.6.: Methodology for building a careflow management system
Quaglini[71]

2.3.4. Modelling care plans

In contrast to clinical guidelines, there have been few attempts to construct a formal model for representing care plans in a computer-interpretable way, although the Asbru formalism[70], with its plan-based, intention-oriented semantics, might be considered to offer a starting point. All clinical activities should be carried out for a specific reason: usually either to achieve, maintain or avoid a situation or patient state. However, clinical guidelines often fail to make explicit the clinical goals underpinning their recommendations[10].

Figure 2.7 provides a high-level model of goal-based clinical decision making derived from the goal properties proposed by Hashmi et al. (cited in [10]). Drawing on AI research on goal-directed behaviour, Fox et al.[10] proposed a process model that evaluates a set of goal properties, summarised in Figure 2.8. Based on the PROforma clinical process modelling language[28], they then proposed a high-level ontology of goal classes and attributes. This consists of two core goal classes: *knowledge goals* (acquisition of information, deciding between alternative hypotheses) and *action goals* (achieve some state, enact tasks). The ontology class hierarchy is summarised in Figure 2.9. Goal descriptions typically consist of an antecedent entry point (e.g. '*if the patient is mobile and can self-care*'), verb phrase–noun phrase action pairs (e.g. '[discharge the patient]$_{\mathrm{VP}}$ to [the intermediate care team]$_{\mathrm{NP}}$'), a temporal constraint ('within [3 days]$_{\mathrm{Duration}}$'), requirement (e.g. optional or obligatory), and rationale (e.g. from a National Service Framework or NICE guidelines). Potentially, parsing out these goal statements from protocol or guideline documents and the free text of patient notes may be possible with information extraction techniques (see Chapters 4 and 5).

However, a fully developed syntax and semantic for formally specifying clinical goals that affords computational reasoning and inference remains to be developed. Recently, Grando et al.[87] have added to PROforma a stateful *Goal* task-type (e.g. with states *dormant*, *in progress*, *suspended*, *completed*) that models the intention (*achieve*, *maintain*, *prevent*) of the transition from an initial patient state to a target state shown in Figure 2.7.

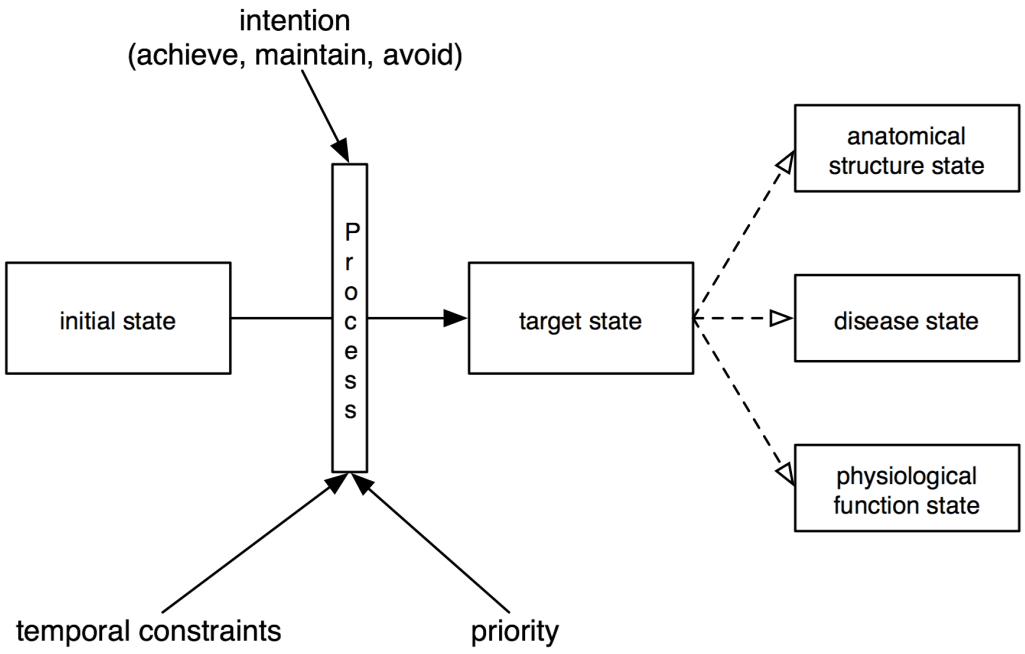


Figure 2.7.: Patient state transitions in goal-based clinical reasoning
Source: derived from descriptions from Hashmi et al. in [10]

2.3.5. Modelling care pathways

As discussed in Section 2.1.5, while there are a number of views on what constitutes a care pathway and how it should be developed, there is some agreement that there should be four components[19][88] [89]:

1. a process map, or workflow, determining the sequence of steps and activities that should be performed, decision points within the process, and the roles assigned to have responsibility for each step;
2. a timeline specifying when each of the activities in the process map should be performed;
3. evidence-based outcome measures, milestones, guidelines and protocols;
4. a 'variance record' i.e. method for documenting and recording where deviations from the planned pathway have occurred. A variance can include:
   - an activity performed by a different role than planned;
   - an activity performed later or earlier than planned;

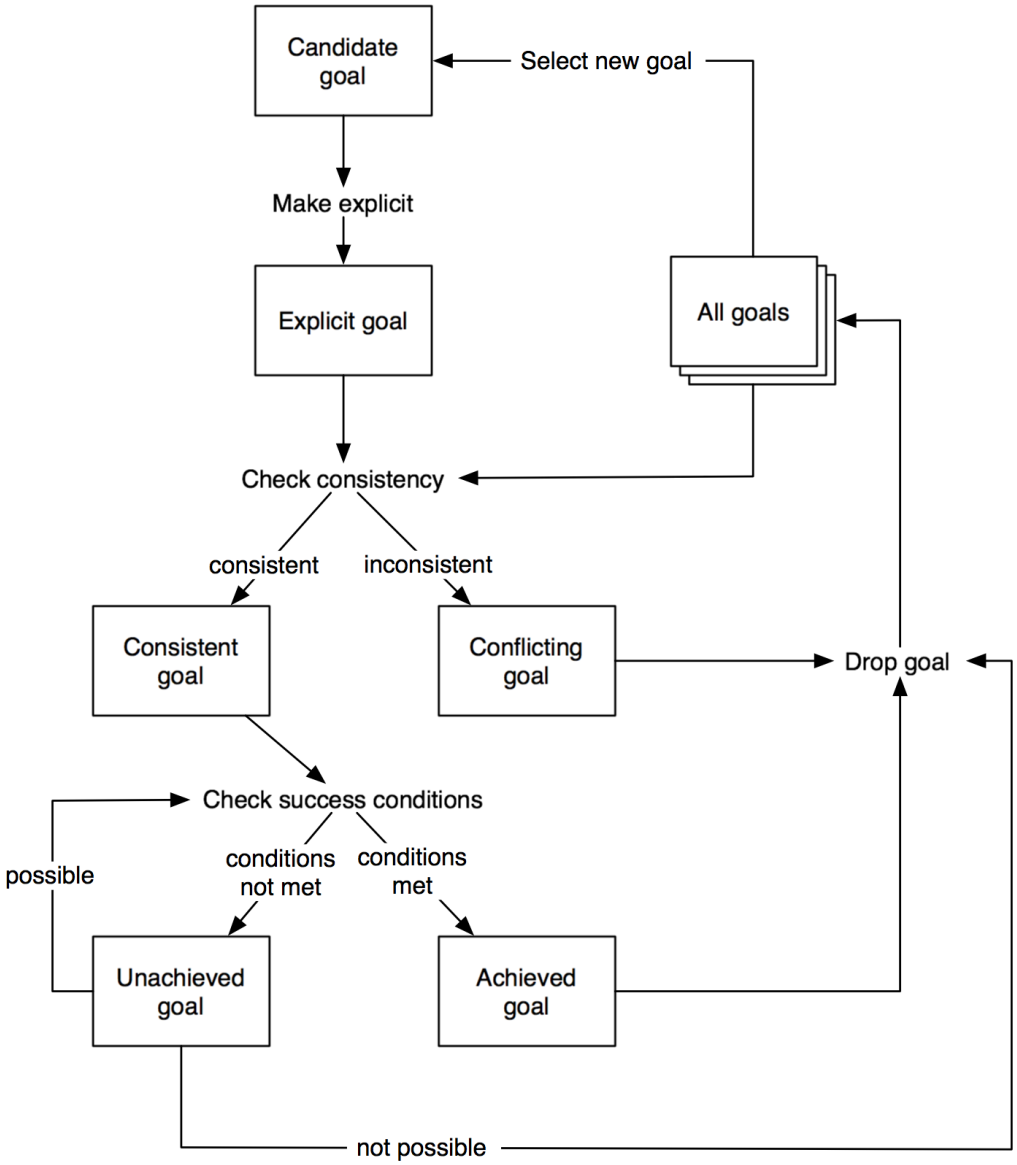


Figure 2.8.: Goal process model
Source: derived from descriptions in Fox et al.[10]

- an activity repeated more or less often than planned;
- an activity being omitted;
- additional, unplanned activities being performed.

de Luc[89] defined electronic care pathways ('e-pathways') as systematically developed, computerised care pathways that describe: (1) the clinical data sets used (representation of declarative knowledge); (2) the on-screen forms and user interface elements required;

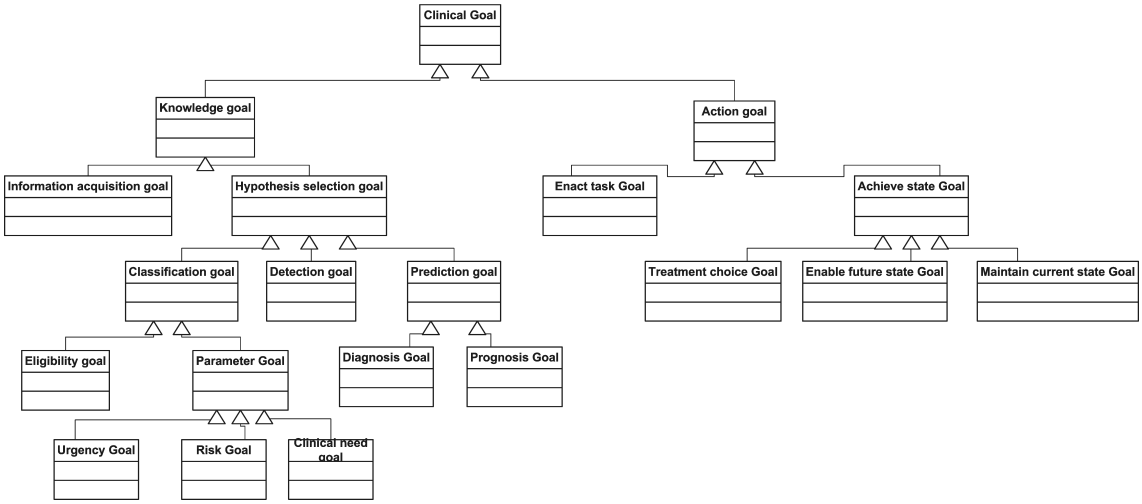


Figure 2.9.: High level ontology of goal classes
Source: derived from descriptions in Fox et al.[10]

(3) the formal model of the roles, tasks, sequencing, and business rules of clinical workflow (representation of procedural knowledge); and (4) the messages to be exchanged between the systems that invoke the pathway. Wakamiya and Yamauchi proposed five core requirements for electronic care pathway implementations: recording notes in the EHR, statistics and variance recording, provision of computerised physician order entry, activity checklists, and editable pathway templates[13].

Page and Herbert[90] developed an object-oriented e-pathway conceptual model using the Unified Modelling Language (UML) formalism. Their model distinguishes a 'model' care pathway template (much like a workflow process definition) from the instantiated, in-use pathway (similar to a workflow instance), which has a state (under consideration, in use, ended). Their model represents the pathway goal, patient entry criteria, and individual pathway activities. Each activity has an activity state, roles, patient, and patient state.

de Luc and Todd[91] proposed the concept of 'generic pathways' for representing an idealised patient journey through an idealised service for a given clinical condition, and distinguish these from 'localised pathways' that have been adapted for a specific institution. These ideas are mirrored by the concepts of generic guidelines and localised, consensus guidelines described by the CIG community[92]. In addition, the Page and Her-

bert model[90] shares a number of features of the task-network CIG formalisms discussed in Section 2.3.2. It is not surprising, therefore, that recent computerised care pathway implementations have used these formalisms (see below and Chapter 3).

Literature on care pathway implementations suggests that a combination of systems approaches (see Section 2.2.2) and discrete-event models (see Section 2.2.1) are required to model complex clinical processes. Todd[93] argues for a systems approach to developing care pathways for the following reasons:

- care pathways are multidisciplinary and holistic; they involve many stakeholders and so require a whole-systems view;
- care pathways involve both clinical and patient goals;
- lower level process maps can be developed from higher-level concept maps.

Systems modelling tasks for care pathways include[93]

1. organisational change (organisation-level process mapping, service commissioning);
2. evidence management (best evidence search and appraisal);
3. computerisation:
    - definition of data sets;
    - design of onscreen forms;
    - organisational ontology design (roles, tasks, sequencing, business rules, clinical workflow);
    - message definitions[93].

A SSM approach to developing care pathways has been described for the management of stroke[93] where the output was a paper-based pathway that represented a high-level view of the patient journey: a computerised version was planned but it was not clear how SSM was to be used to translate the paper pathway to electronic form. SSM was used for modelling chronic disease pathways across different clinics[94], where the output was a

series of high-level clinical workflow and information flow diagrams, but no implementation was provided.

Evidently, both system dynamics and SSM are useful for high-level service modelling and defining the organisational and clinical processes that make up a care pathway. However, these then need to be developed into a discrete-event model for modelling clinical processes at the individual patient level. This modelling hierarchy is visualised in Figure 2.10[95].

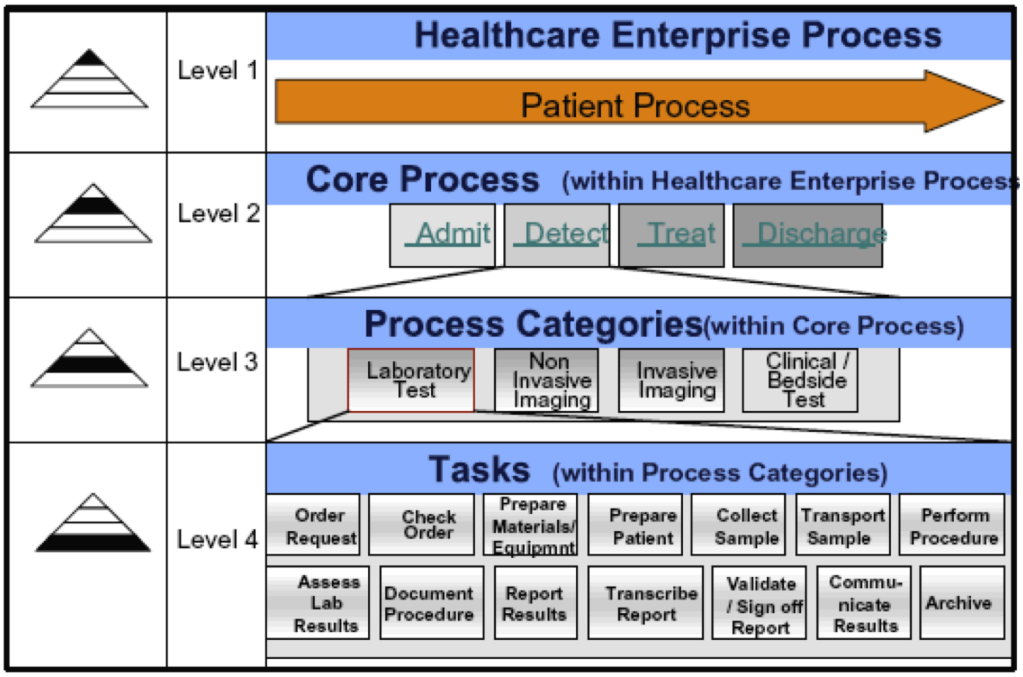


Figure 2.10.: Healthcare service model.
Levels 1 and 2: SD and SSM service model; Level 3: process ontology level; Level 4: discrete event level for clinical and organisational task decision support.
Source: Dang et al.[95]

A precedence diagram method (PDM), based on Gantt and PERT/CPM techniques, was used as the basis for Chu's[36] computerised care pathway implementation. This application used an activity-on-node network, where each node expands to a hierarchical network of composite tasks — a 'multi-level care map' (Figure 2.11). In Chu's PDM model, the finish-to-start relationships of PERT/CPM nodes were augmented with finish-to-finish and start-to-start constraints to allow two tasks to start simultaneously, or to start or finish one after another within a specified lag time. However, these techniques, while useful, may have limitations for modelling clinical processes at the individual patient level,

as they do not provide explicit support for personalisation or dynamic modification[96]. The Asbru computer-interpretable guideline formalism (see Section 2.3.2) has explicit constructs for modelling PERT/CPM/Gantt-type EST, EFT, LST and LFT constraints, making it another potential formalism for modelling care pathways (Figure 2.12).

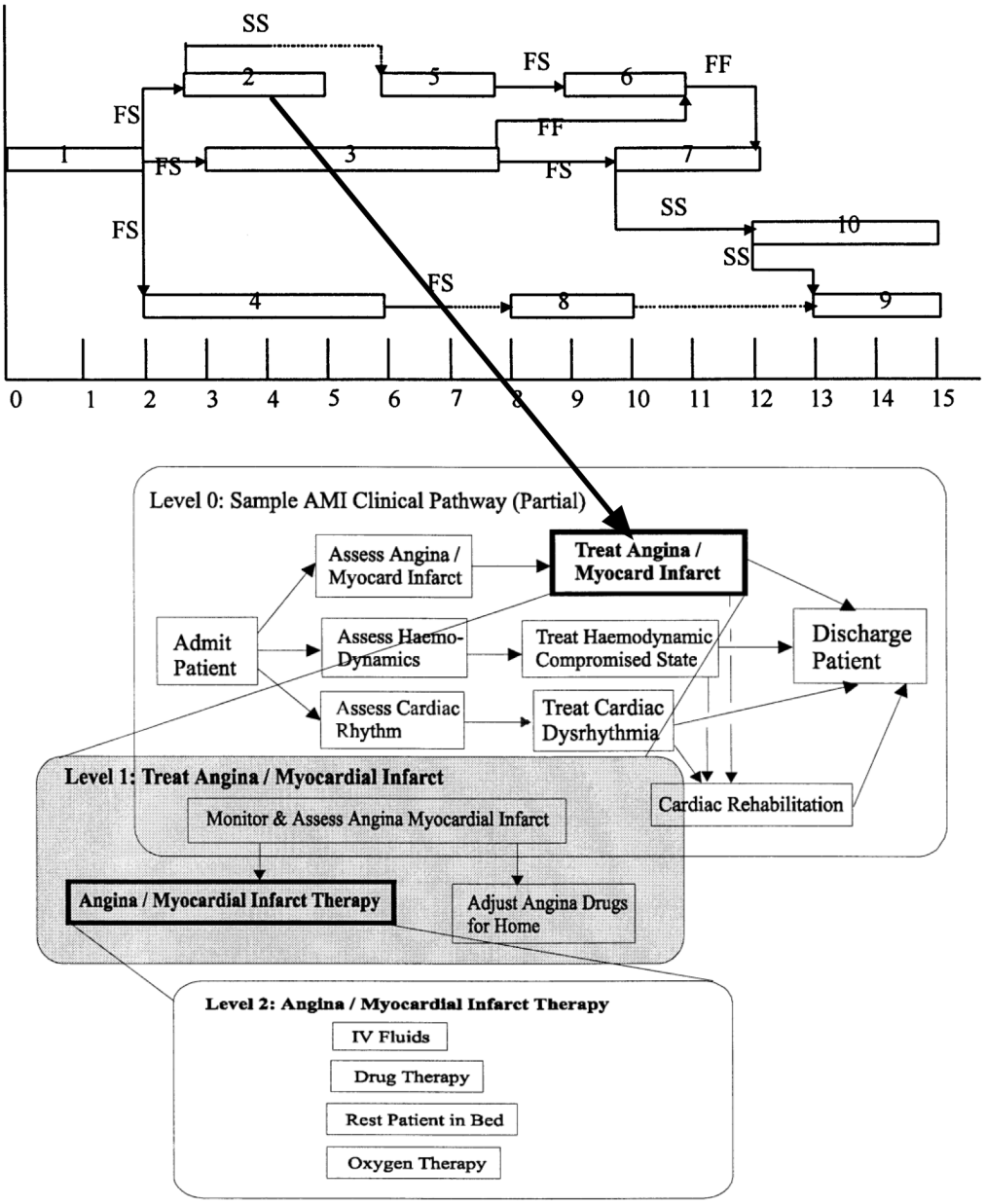


Figure 2.11.: Activity-on-node precedence diagram for representing temporal constraints between care pathway tasks (top) and composite task decomposition (bottom)

Source: Chu et al.[36]

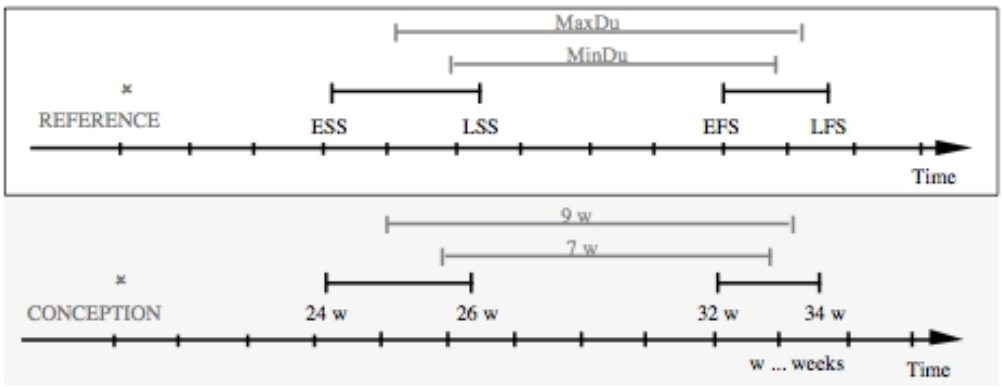


Figure 2.12.: Modelling temporal constraints in processes of care
Source: Shahar et al.[37]

The extended Petri Net model that formed the basis of the 'careflow' framework[71] described in Section 2.3 meets many of the required components of a care pathway described above. Petri Nets were selected as the CIG representation formalism for careflows as guideline and workflow properties could be verified and validated using standard tools available, allowing simulation runs to be performed before enacting the careflow in a live setting. The use of sub-nets to represent composite tasks provides a more formal representation of the multi-level care map approach of Chu[97] (Figure 2.13).

Given the similarities between the task-network guideline models (see Section 2.3.2)[65][82], it is perhaps not surprising that other CIG formalisms can also be expressed as Petri Nets. Grando et al.[29] demonstrated that the PROforma formalism could be mapped to an equivalent PN representation. Combi et al.[27] argued that Petri Nets are a natural candidate formalism to cope with clinical guideline semantics, as they are explicitly geared towards the representation of processes, and are equipped with powerful verification mechanisms. They presented examples using the GLARE CIG formalism.

The CREDO project[98] uses the PROforma formalism to model the 'triple assessment' pathway into a decision support system for the assessment and management of breast cancer. The care pathway was based on assessment and treatment recommendations from a number of clinical guidelines. The system comprised a visual, knowledge formalisation tool, and an execution engine for running the pathway in a Web browser (Figure 2.14). The results of a limited evaluation using simulated cases were reported. While adherence

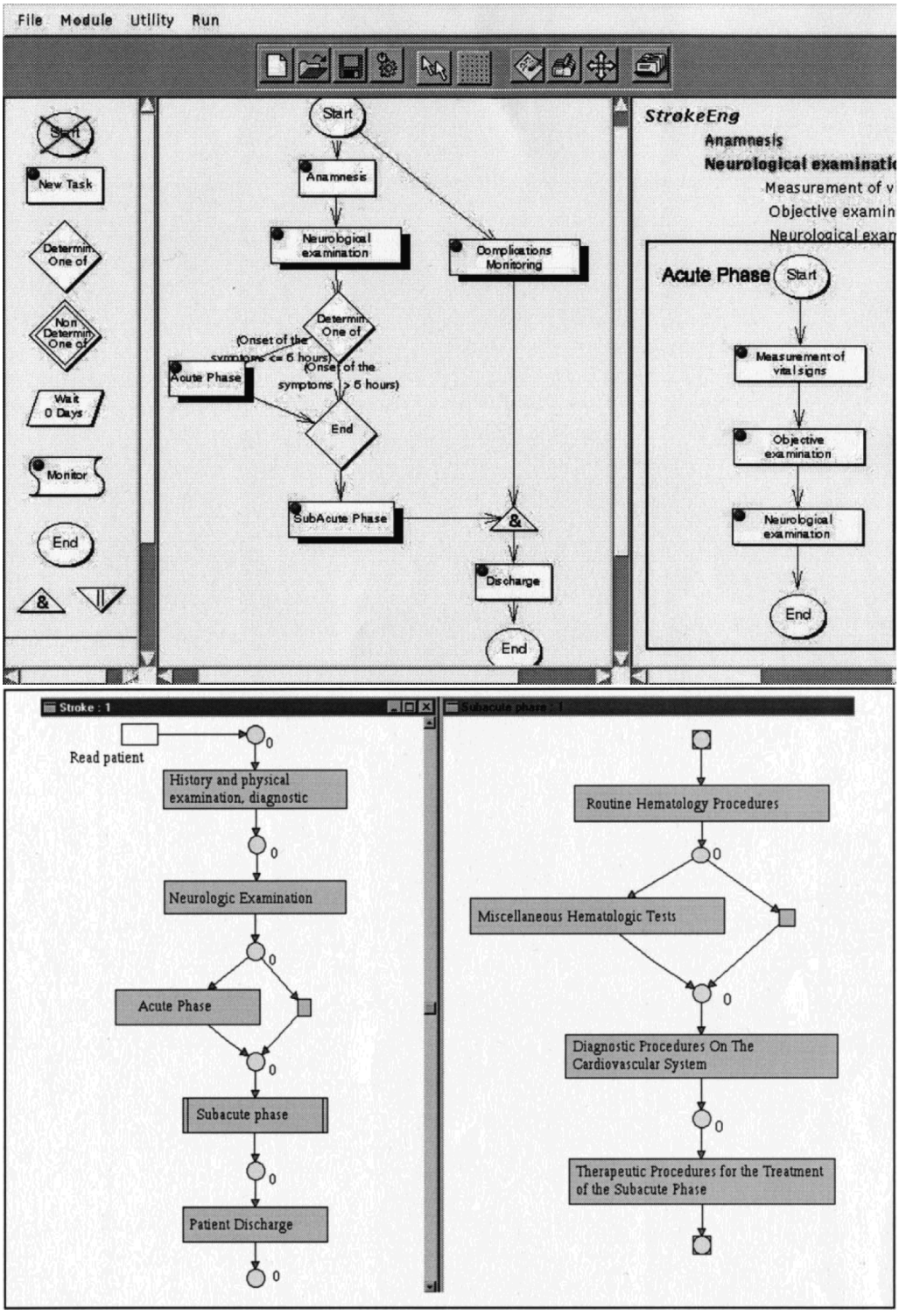
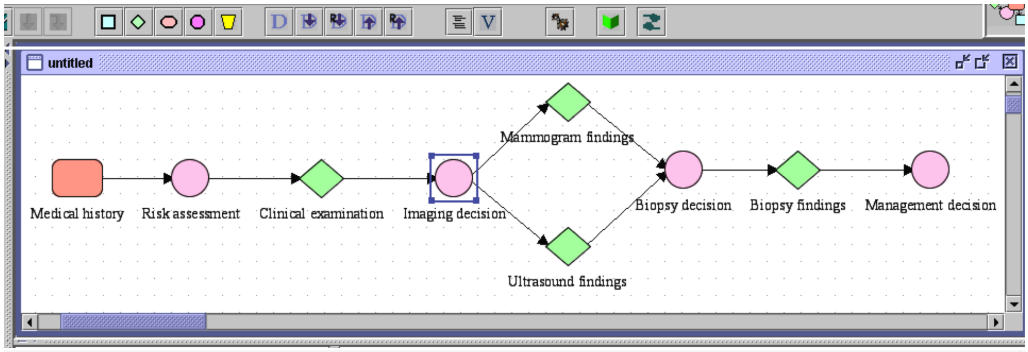


Figure 2.13.: Graphical GUIDE editor for modelling stroke guideline (top) and formal PN representation (bottom)
Source: Quaglini et al.[71]

to best practice guidelines was improved, the strict workflow and task sequencing imposed by the system was seen as a potential hindrance if used in live clinical settings — findings common to other studies on clinical workflow[3].



Decision: Select the relevant intervention to link to arguments for and against

Candidates

- ✓ ☑ **Do an ultrasound of the affected area** ⊟
  - ⊕ The patient has a palpable breast lump More...

  The use of US as an adjunct to mammography resulted in an increase in diagnostic accuracy (D). It is more useful in women with dense breasts. [ACR Practice guidelines for the performance of a breast ultrasound examination]

  References

  1. Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, van Voorthuisen AE. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. Radiology. 1999 Nov;213(2):413-22.

- ✗ ☐ **Neither** ⊞
- ✗ ☐ **Do a mammogram of both breasts** ⊟
  - ⊕ The patient has a palpable breast lump More...
  - ⊕ The patient has some other findings which suggest possible breast primary More...
  - ⊖ The patient is pregnant or possibly pregnant More...

commit

Figure 2.14.: Design (top) and web-based enactment (bottom) of the triple assessment care pathway

Source: Patkar and Fox[98]

2.4. Summary

This chapter has provided an overview of various approaches to modelling processes of care via clinical guidelines, care pathways and workflows in the provision of decision support. The common aim of each is to improve clinical practice via better communication, documentation and adherence to guidelines. Evidence from systematic reviews suggest that this is the case, if such systems are integrated with the EHR, day-to-day working practices and provide support at the point of care. However, these reviews do not distinguish between systems that provide support for individual clinical decisions, and those that manage care processes that extend over time[96].

Individually, the various models and formalisms discussed are all complementary[96], and as we have seen, can potentially be used to model both clinical guidelines and pathways, but the question of how to combine them and integrate them with the patient record and with clinical workflow has rarely been addressed. In the following Chapter 3, we address this question, by presenting a synthesis of the findings of a systematic review of the literature on how these modelling approaches have been implemented within health information systems.

3. Implementing process-oriented health information systems: review and meta-synthesis

3.1. Introduction[1]

In Chapter 2 we noted that computerised care pathways, guidelines and workflow-based decision support systems aim to improve clinical practice and to enable and enhance clinical workflow. In practice, the relationship between clinical workflow and health information systems is more complex. Over 12 years ago, Schneider[2] noted that:

> 'Implementation of clinical information systems (CIS) still requires major changes in workflow... mastering ... how technology is integrated into the clinical practice of medicine, continues to be the key success factor for producing a usable solution'.

The types of workflow changes that clinical information systems tend to impose on the user have been criticised, such as increased cognitive load and inflexible serialisation of task sequences previously done in parallel[3], and a narrow focus on data entry and fragmentation of tasks and roles previously performed collaboratively[4]. Approaches that attempt to square this circle – implementing models of best-practice clinical decision-making and local organisational processes, while integrating with actual clinical workflow – are the subjects of this chapter.

[1] This chapter has been published in an abbreviated form as 'Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems'[1] with some additional material in Results, Discussion and Conclusion

Song et al.[5] defined 'computer-aided healthcare workflows' as the integration of guidelines and protocols with a health information system (HIS) such as an EHR. We extend this idea and define a *process-oriented health information system* as one that formally models guidelines, workflows, or care pathways and provides support for clinical decisions that extend over time. In this chapter, we carry out a systematic review of the literature on process-oriented health information systems to identify the implementation challenges, and the features critical to success. Of particular interest are the steps involved in developing a clinical and organisational workflow model from text-based clinical guidelines, and what cognitive, informatics, technical and organisational resources are required for implementation and integration with day-to-day workflow.

The objectives of this review are to:

1. Provide an account of the processes, resources and challenges involved in developing a process-oriented HIS
2. Compare implementations in terms of knowledge engineering and system architectures
3. Examine the human and organisational factors involved in the development of the system

Review questions include:

1. How are formal knowledge models used in process-oriented HIS created from textual resources?
2. How are HIS integrated with the EHR?
3. How are such systems integrated into clinical workflow?
4. How are such systems made available at the point of care?

3.2. Methods

This review considers a number of phenomenological questions about a process. When one wants to explore a phenomenon about which little is known, in order to gain greater

understanding and develop hypotheses to explain the phenomenon, qualitative methods are an appropriate choice[6]. Therefore we reviewed the literature from this perspective, by treating each paper as a textual narrative from which to extract and categorise the underlying themes that describe the studies as a whole.

The need to include qualitative data in systematic reviews, and the need to undertake reviews of qualitative research in a systematic way, has been recognised for some time[7][8][9], although there are three main practical problems involved:

1. How to select qualitative studies for review.
2. How to appraise the quality of qualitative studies.
3. How to produce a meta-analysis of qualitative evidence[9]

On the first problem, selection, Evans et al.[7] propose selection criteria based on similarity of participants, study focus and themes, and description of qualitative method. Regarding the second problem, appraisal, although there seems to be little agreement on criteria for appraising qualitative literature, there are core principles:[7]

1. clear description of method;
2. clear description of participants
3. evidence of a data trail, so that theme and category labels, which may differ across studies, can be mapped to a unified set during synthesis[7]

These principles informed the basis of our selection criteria. Finally, for meta-analysis of qualitative data, Dixon-Woods[9] suggests identifying variables within qualitative data and weighting them according to the strength of evidence. However, for qualitative data, it is also important to create a *meta-synthesis*: an interpretative analysis of the themes and categories from a representative sample of studies[7]. In this review, we used both approaches. Within the qualitative research field, study heterogeneity is accepted[7], so differences were compared and contrasted, and areas of commonality identified through a process of iterative, comparative analysis.

3.2.1. Search strategy and inclusion criteria

Searches were performed using ScienceDirect, Web of Science, PubMed, and the specialist health informatics OpenClinical web resource. Articles in English published since 1995 were considered in order to analyse how implementation processes have evolved over time. The broad search concepts of health information systems, computerisation, modelling, workflow, pathways, and guidelines were combined into search statements specific to each database queried.

The following broad search concepts were used to query ScienceDirect and Web of Science:

Concept 1: computer systems (systems OR electronic OR computer*) AND
Concept 2: healthcare (health* OR clinical OR care OR medical) AND
Concept 3: guidelines and workflows (pathway OR workflow OR careflow OR guideline)

These three concepts were combined to perform a title search on ScienceDirect and Web of Science:

TITLE ((systems OR electronic OR computer*) AND (health* OR clinical OR care OR medical) AND (pathway OR workflow OR careflow OR guideline))

The following all-fields search statement was performed in ScienceDirect:

ALL (workflow pathways plans guidelines)

The following search statements were executed on PubMed and the results combined:

- (electronic OR computer-interpretable OR computerized OR computerised) AND ((care OR clinical) pathway)
- modelling AND ((clinical guideline) OR ((care OR clinical) pathway) OR workflow)
- workflow AND ((care OR clinical) pathway)
- (clinical guideline) AND ((care OR clinical) pathway)

An initial screening of titles and abstracts excluded opinion pieces, editorials, letters, posters, studies related to non-computerised care pathways, and studies about other types of pathway, for example, biochemical, neural, or motor pathways. Papers on 'patient flows', 'pathways to care' and 'commissioning pathways' were also excluded at this stage as these focus on the larger goal of strategic planning rather than clinical workflow and decision making at the individual patient level. Reviews of CIG and workflow models (such as those described in Chapter 2) were selected as background material, and were used as a source of additional citations.

Full text articles were screened and included if they met the three inclusion criteria: (1) the study addressed the modelling process for the computerisation of clinical workflow, clinical guidelines, or care pathways within the context of a HIS; (2) the outcome was the exposition of a new methodology, knowledge model, framework, system implementation, or system architecture that instantiated the process under study; and (3) there was an evaluation, even if this was only formative and descriptive.

We generally excluded trials, retrospective EHR data analyses and systematic reviews of the clinical impact (on rates of medical error, adherence to guidelines, length of stay, costs, etc.) of care pathways or solely of the relative effects of 'computer generated guideline recommendations' or 'computerised guidelines' vs. paper guidelines. First, such studies are addressing different questions to our review. Second, this area has already been well-covered in the CDSS and care pathway literature (e.g. Kawamoto et al.[10], Rotter et al.[11]). Third, such studies tend not to discuss the informatics process of 'computerising' the guideline (e.g. Eccles et al.[12]) or how the computerised guidelines are validated with respect to the original text guideline (e.g. Rood et al.[13]). A 'computerised' guideline might simply be the full-text of the guideline in HTML or other electronic format (the clinician still has to read it). However, if such a study also addressed the aims and objectives of this review, it was included.

3.2.2. Data collection and quality assessment

Following Evans and Pearson,[7] a data collection form was created in Microsoft Excel to identify papers for review. The quality of each was judged using criteria from Burns[14] and Greenhalgh[6] such as a clearly formulated question, rationale for and description of setting and participants, methodological, theoretical, and analytical rigor, data audit trail, and justification of conclusions.

Information for each of these criteria from each study was entered into the data collection spreadsheet[2]. Not all criteria were relevant for each paper (e.g., model formulations and system architectures may not have any participants or data audit trail). Papers that could not meet the criteria were discarded.

3.2.3. Data abstraction and thematic analysis

Thematic analysis was carried out using an approach informed by qualitative concept analysis, in which research aims are defined in advance, and categories are brought to the material and continually refined against it, with the goal of reducing the material[15]. This was guided by the three-stage approach discussed in Miles and Huberman[16]: (1) initial, descriptive coding, developing toward (2) more interpretative coding (high-level concepts that encompass the descriptive coding performed in step 1) as knowledge of the phenomenon under study increases; and (3) pattern coding (emerging themes) toward the end of the analysis in which themes are developed that seek to explain and make causal links in the phenomenon.

Challenges identified by Song et al.[5] were used to help develop the initial working list of descriptive codes with which to annotate the data (step 1 described above). Briefly, these were:

- data collection and normalisation
- workflow integration
- legal/regulatory/safety issues

[2]The complete spreadsheet is too large to reproduce here; it is available on the CD that accompanies this thesis.

- usability
- data/workflow visualisation
- adaptability
- flexibility
- maintenance
- systems integration
- validation and verification
- workflow formalisation

The list of codes was refined and enhanced as new themes emerged from the literature during analysis (step 2). Examples of emerging themes used to code the literature are shown in the tag cloud in Figure 3.1. The text size of each term in the tag cloud is proportional to its frequency in the corpus (theme frequency shown in parentheses after each term).

Figure 3.1.: Initial coding themes that emerged from thematic analysis (Theme frequency shown in parentheses)

The final set of pattern codes was used to thematically annotate each paper in the review (step 3). Up to five variables that reflected the study's key concerns, results, and conclusions, were assigned to each study – these were the 'challenge theme' variables, that is, factors that need to be addressed when developing a system.

RefViz[17] is a tool for clustering bibliographic references for visualisation and analysis. We created a custom reference file in ISI ResearchSoft RIS format[18], where each entry contains title, year, author, and challenge theme variables for each paper, as in the following example:

```
TY  - JOUR
ID  - J631
T1  - Embedding Oncologic Protocols into the Provision of Care:
      The Oncocure Project
A1  - Eccher
Y1  - 2009
N2  - System-architecture, Separation-of-concerns, Data-mapping,
      Process-modelling
```

This reference file, containing entries for all selected studies, was imported into RefViz for analysis. RefViz applies standard mathematical clustering algorithms to partition the data set into concept-based groups of similar papers based on the co-occurrence of themes between papers. RefViz's Galaxy view performs principal component analysis (PCA) in which a larger set of possibly correlated variables are transformed into a smaller, more fundamental set of independent variables[19]. The co-occurrence and clustering of the challenge theme variables arising from the thematic analysis were explored using PCA in RefViz, in order to see if the set of variables could be transformed into a smaller number of principal components that further summarise the studies and from which an integrative, conceptual model of the implementation process could be developed.

3.3. Results

From 1308 screened citations, 200 full text articles were retrieved, and 108 met the inclusion and quality criteria for detailed review. The selection process is shown in Figure 3.2.

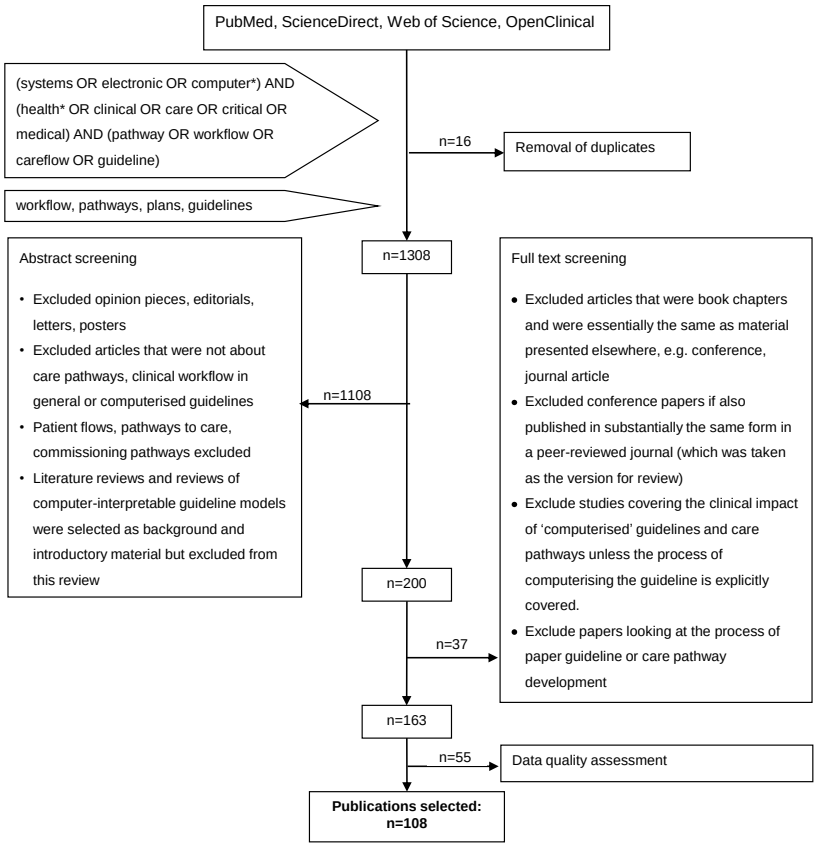


Figure 3.2.: Screening flowchart
Source: Gooch and Roudsari[1]

3.3.1. Publication date distribution

In the more general area of clinical decision support, Greenes[20] identified five distinct phases of research interest between 1960 and the present:

1960–1985: 'a long infatuation' – initial research enthusiasm;

1985–1998: 'a troubled courtship' – implementations showing benefit but limited dissemination outside the academic environment;

1998-2003: 'renewed passions' – knowledge explosion, safety and quality agenda;

2003–: 'long lasting relationship' — full-systems implementations of EHR, CPOE, e-prescribing; and improved understandings of requirements;

2004–: 'a new party to the relationship' – recognising knowledge management as a necessary infrastructure.

It may be possible to identify similar trends in the modelling and implementation of process-oriented health information systems in the selected studies, which are explored in the remainder of this review. Figure 3.3 shows the publication date distribution of the search results after abstract screening but prior to full-text screening ($n = 200$). There appears to be an overall trend towards increased interest in computational modelling and execution of clinical guidelines, workflows and care pathways, although this trend may not be significant ($R^2 = 0.3469$).

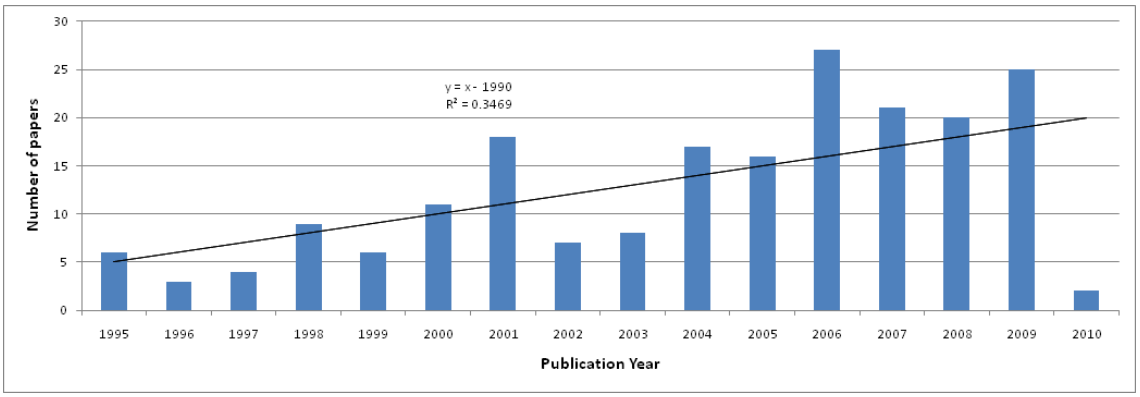


Figure 3.3.: Publication year distribution of selected studies

3.3.2. Characteristics of selected publications

The review identified 79 journal articles and 29 conference proceedings papers. Fifty-seven (53%) studies were conducted within a non-clinical academic or commercial research and development environment. The remainder took place within university teaching hospitals and medical centres ($n = 16$, 15%), outpatient clinics ($n = 8$, 7%), and general hospitals, stroke units, or emergency or ICU departments ($n = 27$, 25%).

Methods used by selected studies ranged from qualitative research involving usability evaluations ($n = 1$) or questionnaires, interviews, and observational studies ($n = 20$), to formal methods papers ($n = 26$), model formulations ($n = 26$), system case studies

($n = 20$), prototype implementations ($n = 33$), and system architectures ($n = 26$). These categories were not mutually exclusive; a number of studies had multiple objectives: for example combining model formulation, prototype implementation, and system architecture.

Eight distinct knowledge model types were identified in the publications. Fifty-four publications (50%) focused on providing details of system architecture or system prototype implementation. Forty-four (41%) studies had evaluation results reported in the form of interviews, questionnaires, and observational case studies where the study size was quantified. The remaining studies reported informal evaluation in terms of the features of the model or method, or overall benefits of the system implemented.

3.3.3. Challenges in implementing process-oriented systems

The final set of the 25 challenge theme variables and their descriptions, derived from thematic analysis of the 108 papers, are shown in Table 3.1. The association between themes was explored using the Galaxy and Matrix views within RefViz. The weight of each theme within each cluster is calculated by RefViz's implementation of PCA and indicates the strength of association between the theme and the cluster, on a scale from -1 (strongest negative association) through 0 (no association) to +1 (strongest positive association). RefViz identified ten clusters (see Appendix A for the complete matrix of association scores), from which the concept map shown in Figure 3.4 was developed.

In Figure 3.4, each cluster is shown as a circle, where the radius of the circle is proportional to the number of papers in the cluster. Only the positively associated themes (i.e., with non-zero or non-negative weights) are shown, and the thickness of the line is proportional to the strength of association between the cluster and the theme. Table 3.2 provides a description of each challenge theme cluster, where the numeric group identifier relates to each cluster in Figure 3.4.

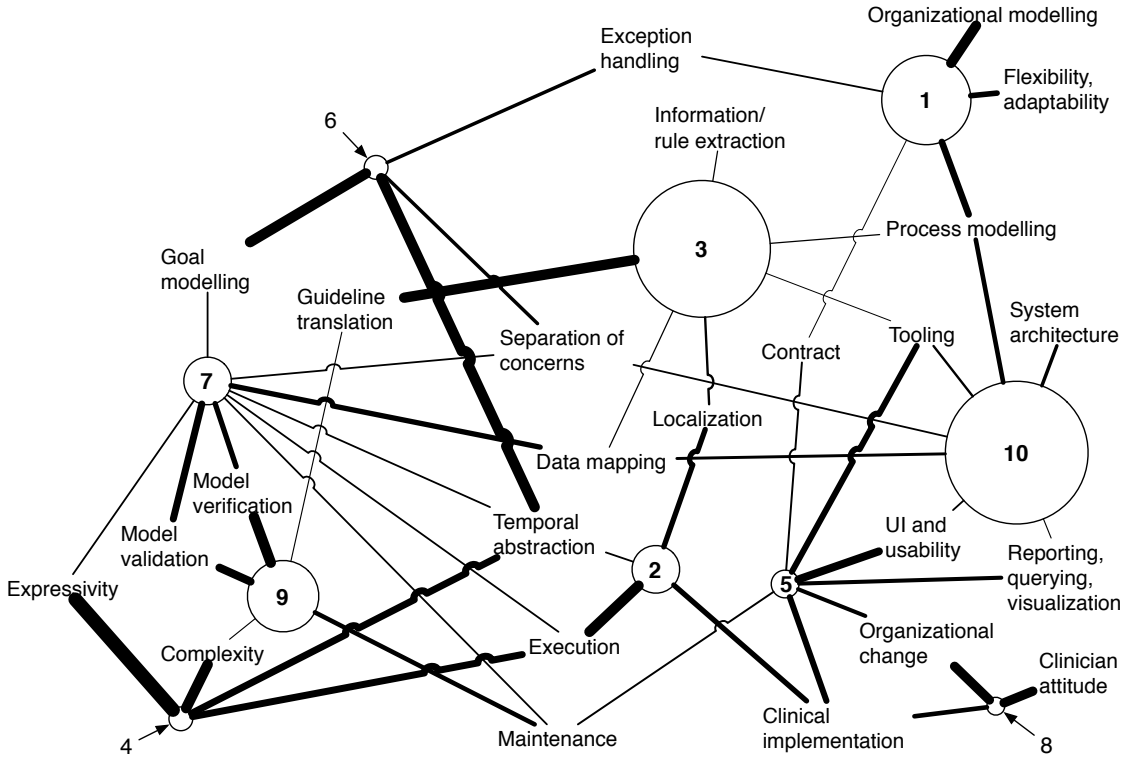


Figure 3.4.: Concept map derived from RefViz Galaxy and Matrix analysis, showing association between study clusters and the 'challenge theme' variables.
The radius of each circle is proportional to the number of studies in the cluster; the thickness of the line between cluster and theme is proportional to the strength of association between the cluster and the theme.
Source: Gooch and Roudsari[1]

Table 3.1.: Challenge themes: 25 variables identified from initial thematic analysis

Variable	**Description**
Clinical implementation	Implementing the model into a usable system that is congruent with individual and collaborative clinical workflow in a live, clinical environment
Clinician attitude	Beliefs in own self-efficacy, and relevance and quality of guidelines and pathways to clinical practice

Continued on next page

Table 3.1 – continued from previous page

Variable	Description
Complexity	Ability to evaluate and check the model with reasonable run-time behaviour (e.g. polynomial time) in real-world scenarios
Data mapping	Mapping electronic health record (EHR) data to procedural tasks in the guideline or pathway; mapping guideline concepts to terminologies
Contract	There is an implied 'contract' between system workflow documentation and treatment process; incorrect or unexpected system use, staff miscommunication, or model/implementation constraints have the potential to cause divergence between system records and actual treatment (e.g. ticked action not actually being performed)
Exception handling	Ability to handle unplanned deviations from the pathway or guideline (variance)
Execution	Executing the guideline or pathway model within the EHR; semantic interoperability
Expressivity	The need to adequately represent complex clinical information, rules, and exceptions in a formal model
Flexibility and adaptability	Adapting the pathway at run-time to individual patient (variance); handling incomplete or ambiguous patient data
Goal modelling	Modelling clinical and organisational processes is insufficient: the intention for each task needs to be explicit
Guideline translation	Guidelines are ambiguous and cannot easily be translated into logic rules; contain implicit knowledge that is incompletely specified

Continued on next page

Table 3.1 – continued from previous page

Variable	Description
Information/rule extraction	Ability to automatically extract clinical knowledge and rules from guideline text
Localisation	Adapting the pathway to local needs (consensus and collaboration). Domain experts creating shareable guidelines must agree on meaning and interpretation of the guideline
Maintenance	Need to keep guideline, pathway, and workflow model up to date with latest evidence or changes in clinical workflow
Model validation	Validation of encoded model against clinical relevance and expected results for the specific patient; explanation of reasoning
Model verification	Internal consistency of the model, well formedness, proofs of properties
Organisational change	Existing clinical workflow may need to be adapted in order to successfully implement the system. Staff buy-in, training, and workflow needs; changes of role (e.g., increased data entry at point of care)
Organisational modelling	Need to model organisational workflow as well as medical knowledge; includes role-based access and security
Process modelling	Creating a computer-interpretable model of clinical processes from guidelines and local clinical knowledge
Reporting, querying, and visualisation	Getting access to the data held in the system for reporting, statistics, visualisation
Separation of concerns	Separation of medical knowledge from workflow knowledge that can be integrated into a combined clinical and organisational process model at run-time

Continued on next page

Table 3.1 – continued from previous page

Variable	Description
System architecture	Selection of a suitable system architecture congruent with clinical workflow and organizational needs: for example, client-server, service-oriented architecture (SOA), semantic web, transport layer security, authentication, role-based access
Temporal abstraction	How to model temporal constraints and periodicity in guidelines and pathways
Tooling	Creation of easy to use tools to model guidelines, workflows, and pathways
User interface and usability	Accessing the data and guideline/pathway in an easy to use, easy to navigate way; data entry

Table 3.2.: Description of the challenge theme clusters shown in the concept map of Figure 3.4

ID	Studies in the cluster	Cluster description
10	24 Studies[21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41][42] [43] [44]	Creating a procedural, clinical process model aided by knowledge acquisition tools and supported by the system architecture; mapping declarative concepts between a local electronic health record (EHR) or 'virtual medical record' model and the process model; user interface (UI) and usability design congruent with the model; separation of organisational, medical, and UI models

Continued on next page

Table 3.2 – continued from previous page

ID	Studies in the cluster	Cluster description
3	23 Studies[45] [46] [47] [48] [49] [50] [51] [52] [53] [54] [55] [56] [57] [58] [59] [60] [61] [62] [63] [64] [65] [66] [67]	Collaborative process between informaticians and domain experts of translating implicit, procedural knowledge into computable rules; extracting declarative and procedural knowledge into a process model; localisation of the guideline/pathway for a specific institution and mapping to the local EHR
1	15 Studies[68] [69] [70] [71] [4] [72] [73] [74] [75] [76] [77] [78] [79] [80] [81]	Integration of clinical and organisational processes with regard to institution-specific clinical workflow and preferences; handling workflow exceptions (adaptive organisational workflow); bindings/congruence of enacted workflow with documented clinical processes
9	12 Studies[82] [83] [84] [85] [86] [87] [73] [88] [89] [90] [91] [92]	Verification and validity of the clinical process model; formal proofs; model-driven update and maintenance of the knowledge base
7	8 Studies[46] [93] [94] [95] [96] [97] [98] [99]	Clinical validity of EHR–guideline concept mappings; verification of rule-set completeness and consistency; verification and validation of temporal constraints and run-time execution
2	8 Studies[100] [101] [102] [103] [104] [105] [106] [107]	Enactment of the model within local EHR/health information systems (HIS); handling clinician judgment, task sequencing, and temporal constraints, exceptions, variance (adaptive clinical workflow)

Continued on next page

Table 3.2 – continued from previous page

ID	Studies in the cluster	Cluster description
5	7 Studies[108] [47] [109] [110] [111] [112] [113]	Addressing usability barriers to implementation of a computerised guideline or pathway; integration with clinical and organisational workflow; development of new tools to support clinical workflow; modification of existing workflow to fit computerised workflow; reporting workflow/pathway statistics, and exceptions
6	4 Studies[114] [115] [116] [117]	Formal modelling of clinical goals and their temporal constraints; separation of clinical and organisational goals; allowance for unplanned run-time deviations in the model
4	4 Studies[118] [119] [120] [121]	Handling of complex temporal expressions within the pathway that provides adequate abstraction while remaining computable (trade-off between expressivity and complexity)
8	3 Studies[122] [123] [124]	Overcoming the organisational and individual barriers to implementation of a computerised workflow, guideline, or pathway; need for both computerised and real workflow to adapt to each other

3.3.4. Approaches to implementing process-oriented systems

Electronic health record integration

Twenty-six studies considered the problem of how to integrate a clinical process model with data in the EHR. Of these 26 studies, only three[26][99][41] were part of a system implementation within a clinical environment; the remainder were data modelling and/or integration studies within an academic institution. In terms of approach, the studies can

be split into three categories:

- Studies that advocated the use of the same underlying data model for both the guideline or pathway knowledge model and the EHR, using models such as the HL7 Reference Information Model (RIM), Unified Service Action Model (USAM), or openEHR[21][36][38][40]
- Studies that attempted to map guideline or pathway knowledge model concepts to data items within the EHR via guideline expression languages (e.g., GELLO);[25] the use of a 'virtual medical record' (VMR);[52][103][37][41][64] standardised vocabulary resources such as the Unified Medical Language System (UMLS) and SNOMED CT;[52][25][31][32][67]; a 'middleware' mapping ontology layer;[46][125]; or manually, on a system-specific basis,[26][66] or via a translation table[101]
- Studies that recognised the need for EHR integration, but did not implement it[50].

Clinical workflow integration and point-of-care use

Studies that considered the use of guidelines and pathways at the point of care can be divided into model formulations and practical implementations of systems. A number of the model formulation studies suggest that the barrier to the accessibility of guidelines or care pathways might be addressed by developing an ontology that integrates organisational and clinical workflow with EHR data requirements[50] [25] [124] [106]; however, these papers do not suggest how such point-of-care execution should be implemented in practice.

We found that implementations of workflow integration with point-of-care use tended to be one of three types:

1. Use of an integrated device for data collection, display, and guideline-based decision support. Examples included the use of ICU bedside monitoring workstations providing real-time data trending, and care plan and test result information[83], the use of mobile devices providing access to clinical guidelines[57], and an emergency triage pathway implemented as a rules-based expert system in a mobile device[34]. Evaluation details for each of these, however, were brief, tending to focus on the

hardware/software infrastructure and non-quantified statements about system accuracy.

2. Use of electronic patient encounter forms that mirror the structure of existing paper forms. Examples included a guideline-based system for reminders and order recommendations,[55] and a care pathway for proximal femoral fracture[33] where guideline-based recommendations were presented as default selections on the form (e.g., automatically ticked checkboxes). Neither appeared to offer pathways tailored to the specific needs of the patient, nor made it clear how computer access would be available at all points of the clinical workflow.

3. Augmented use of paper forms for system input and/or output. Examples included a rules-based system using guidelines encoded in Arden Syntax that used optical character recognition (OCR) to scan paper forms, completed at the bedside, to provide patient-specific, point-of-care recommendations and reminders,[113] and a system that provided a print-out of daily workflow tasks according to the care pathway modelled. The printed task lists could be used at the point of care as a clinical reminder, but patient-specific recommendations or decision support were not provided[109].

Brokel et al.[26] suggested that merely integrating clinical decision and workflow rules within the EHR was insufficient to ensure care pathway recommendations are made at the point of care: this will only happen if clinicians actually use the system while interacting with the patient. If clinicians are entering data into the system post-hoc, then the benefit of point-of-care advice is lost[26]. This theme of a care pathway offering a 'contract' between what is recommended and what treatment is actually recorded was taken up by Lenz et al.[33], who described the development of a care pathway system that, as in [55], used structured data collection forms, and that integrated guideline-based recommendations as default selections on the form (e.g. automatically ticked checkboxes). However, the 'charting by exception' approach, where only deviations from the pathway ('variance') are recorded, led to some unintended effects. The lack of integration with an order-entry system, coupled with the use of default form selections, caused process steps to be

documented that did not actually take place. The system did not clearly differentiate between the act of placing of an order and the documentation that an order had been placed, nor between previous, current and future medications, as orders had to be placed independently of the system using a paper form.

System implementations: knowledge models, software, and architecture

Table 3.3 defines the eight distinct knowledge model types that were identified. In the studies retrieved, formal task-network models, which support the representation of both guideline concepts and workflow patterns, were the most commonly described and implemented.

These models were instantiated in the 54 studies that described a system architecture and prototype implementation (see Appendix B). Eighteen of these (33%) explicitly implemented clinical workflow support via a defined workflow process and/or workflow engine; and 26 (48%) described integration with the EHR, but this appears to be largely limited to conceptual integration – few studies have implemented this in a live, clinical setting[104]. Eleven (20%) described both workflow and EHR integration.

System architectures ranged from standalone desktop[22] [47] [126] [101] [24] [84] [29] [30] [56] [57] [88] [91] [124] and web browser applications[49] [86] [106] [42] to client-server systems [109] [69] [111] [96] [83] [99] [34] [77] [113] [62] [107] and distributed, web service applications[68] [70] [50] [53] [119] [27] [72] [103] [74] [75] [36] [39] [41].

Systems (not mutually exclusive) included computerised guideline implementations[47] [109] [49] [50] [100] [126] [101] [96] [53] [24] [83] [125] [84] [119] [86] [102] [29] [99] [55] [103] [104] [34] [57] [88] [89] [113] [62] [36] [39] [41] [121] ($n = 31$), computerised care pathway systems[22] [111] [30] [74] [35] [105] [124] [106] [42] [44] [107] ($n = 11$), integrated guideline and WfMSs[69] [27] [103] [77] [78] [36] [39] [43] ($n = 8$), computerised clinical workflow systems[68] [70] [72] [74] [75] ($n = 5$), and automated guideline formalisation and verification applications[56] [88] [91] ($n = 3$). For the pure guideline-based systems, for the clinical knowledge component there was a general trend from the use of ad hoc, procedural code toward the use of more formal, task-network models. For the care pathway systems,

Table 3.3.: Frequency and description of knowledge model types used by studies

Knowledge model	Description
Document model (5 studies, 1 system implemented[75])	Human readable document with concepts represented in situ, usually preserving the original structure of the source document (Guideline Elements Mode (GEM) or other document-centric extensible mark-up language (XML) schema)
Semantic web (9 studies, 6 systems implemented[125] [119] [72] [103] [35] [124])	Models proposed by the world wide web consortium (W3C) for representing information on the web (web ontology language (OWL) ontologies, Semantic Web Rule Language (SWRL) rules, OWL-S web services)
Formal workflow model (8 studies, 3 systems implemented[69] [27] [74])	Formalised workflow constructs underpinned by a formal mathematical model (Petri Nets, Yet Another Workflow Language (YAWL))
Object model (8 studies, 2 systems implemented[47] [62])	Object-oriented techniques to model collection of hierarchical, interacting classes that represent guideline, workflow, or pathway concepts (Unified Modeling Language (UML), HL7 Reference Information Model (RIM), openEHR)
General task-network model (14 studies, 4 systems implemented[22] [100] [77] [78])	Flowcharts or process maps without formal semantics (Program Evaluation Review Technique/Critical Path Method (PERT/CPM), activity-on-node)
General workflow model (14 studies, 11 systems implemented[68] [70] [72] [33] [75] [77] [78] [105] [39] [43] [44])	General workflow semantics (Business Process Modeling Notation (BPMN), Business Process Execution Language (BPEL))
Block-structured, procedural, logic rules (20 studies, 11 systems implemented[68] [49] [126] [84] [86] [55] [34] [57] [88] [89] [113])	Block-structured, procedural programming languages, and IF...THEN rules (Arden Syntax; decision tables)
Formal task-network model (48 studies, 23 systems implemented[69] [47] [109] [50] [101] [96] [53] [24] [83] [27] [102] [29] [99] [56] [103] [104] [91] [36] [39] [41] [42] [121] [107])	Guideline-based clinical tasks – actions, decisions, queries – that unfold over time, with a formal syntax and semantics (Guideline Interchange Format (GLIF), PROforma, Asbru)

the trend was from the use of informal or unspecified models toward the use of a general workflow model with a task-network or semantic web formalism. Only two of these[33] [124] appeared to meet all the requirements proposed by Wakamiya and Yamauchi[108] (see Chapter 2, section 2.3.5).

A number of studies suggested that integration of the care pathway or guideline with an organisation's clinical workflow and EHR requires a tightly coupled architecture, in which system components rely on knowledge of other components' internal workings to access their data directly, share the same global data, or directly control the operations of other components[99] [52] [24] [83] [74] [34] [113] [43]. Tight integration between system components arguably reduces system portability and interoperability but has the benefit of greater efficiency, as few components to broker communication between modules are required[99].

Others proposed a modular approach to reduce coupling between systems. However, these still tended to be database-centric, tied to specific mapping tables, database engines, or commercial workflow tools[109] [100] [77] [78]. Those that integrated a guideline-based system with an existing EHR typically implemented an 'event listener' that monitors the EHR for new clinical events or data from which opportunities for decision support are identified and invoked[69] [83] [102] [103] [104] [78] [107]. This allows more portability between components but potentially at the expense of inefficient use of network and database resources, due to the overhead of creating, transmitting and translating messages brokered by the event listener or messaging component[99].

Some recent approaches utilise a service-oriented architecture (SOA), where standard messaging interfaces (such as hypertext transfer protocol (HTTP) and simple object access protocol (SOAP)) enable loose coupling between applications[70] [53] [125] [119] [27] [72] [103] [75] [36] [39]. Semantic web-based care pathway architectures[125] [119] [72] [103] [35] augment the SOA approach by allowing dynamic, context-aware composition of workflows from individual web services. These use World Wide Web Consortium (W3C) standards such as OWL-S and SWRL for defining classes of services and resources, and the rules that relate them.

3.3.5. Toward a conceptual implementation model

A conceptual model of the implementation process was developed from the theme clusters shown in Figure 3.4 and Table 3.2, and by referencing each cluster back to the studies from which they were derived. The model is shown in Figure 3.5 and described below.

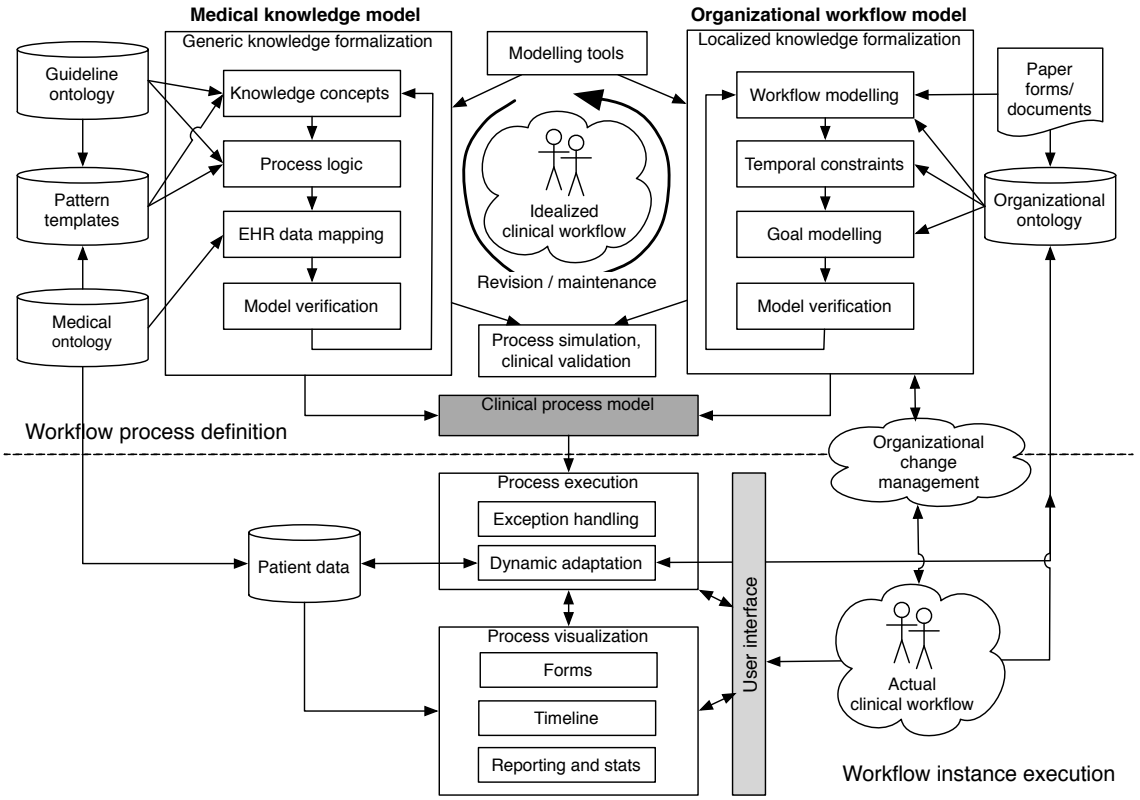


Figure 3.5.: Conceptual model for implementing process-oriented health information systems.

Source: Gooch and Roudsari[1]

Development of a process-oriented HIS is an iterative, collaborative process[45] [49] [50] [23] [52] [26] [98] [54] [30] [60] [79] [41] [66] that involves defining a clinical process model (*shaded in figure*) comprising formalised medical knowledge (usually from guidelines) (*top-left of figure*) and organisational workflow (*top-right of figure*). A graphical knowledge acquisition tool is typically used to assist in this task[68] [69] [71] [50] [95] [100] [84] [27] [86] [29] [33] [104] [36] [39] [107]. The model (typically derived from one or more of the types presented in Table 3.3) represents an idealised view of the knowledge concepts, processes, and rules of clinical workflow required to enact the guideline or pathway, and

tailored to local intervention strategies.

Medical knowledge formalisation typically involves the use of an ontology for the guideline concepts and process logic[69] [114] [48] [93] [50] [100] [52] [125] [85] [25] [119] [98] [72] [102] [34] [77] [35] [106] [64] [42], and a standard medical terminology to map guideline concepts to terms in the EHR data model or VMR[69] [46] [52] [101] [125] [25] [26] [119] [103] [35] [37] [41] [64] [66] [67]. Some researchers have attempted to automate the extraction of clinical rules and process information from guideline statements to ease the process of formalisation, sometimes with a high degree of recall and precision[48] [56] [65]. This has been done via the use of linguistic phrase pattern templates[127] [48], and natural language processing and information extraction pipelines[56] [63]. Such techniques may be useful for facilitating automatic updates to the knowledge base[73].

This generic model needs to be localised to the setting/ institution[69] [52] [103] [89] [66]. This task can be commenced prior to modelling, to create a 'consensus' version of the guideline[93] [50] [126] [53] [104], ready for formalisation, or the encoded, generic model can be shared among institutions, each adapting it according to local needs and data items available in the institution's EHR[52] [96] [27] [89] [106] [66]. Localisation also involves creation of an organisational workflow model, or addition of workflow concepts to the formalised medical knowledge model. Workflow modelling may make use of an organisational ontology[69] [28] [72] [74] [77] to formalise tasks, roles, and treatment goals[69] [114] [93] [116] [32] [117]. Definition of temporal constraints, often not present in the guidelines themselves,[115] is required for activity sequencing and scheduling[100] [118] [28] [115] [90] [65] [120] [121] [117].

Model checking techniques and tools provide formal means of verifying that encoded models are correct and consistent[69] [95] [82] [85] [87] [115] [90] [91], particularly when maintaining and updating them[90]. Simulated runs of the model are used to ensure that the output is clinically valid.[69] [49] [52] [101] [96] [53] [118] [125] [85] [105] [92] [107]

To execute the clinical process model within a HIS, architecture, user-interface design, and mode of delivery need careful consideration in order to be congruent with actual clinical workflow[22] [108] [47] [112] [24] [31] [33] [34] [113]. This can be facilitated via

visualisation of the run-time pathway[24], design of on-screen forms based on the paper forms of a manual care pathway[30] [55] [33], or automatic generation of forms directly from the pathway ontology or process model.[119] The enacted process should allow dynamic adaptation at run-time: this may be manual and clinician-led, where tasks can be skipped, repeated, or new tasks added[68] [71] [96] [75], or may be system-led via reasoning over new knowledge added to the ontology at run-time[72] [35].

Implementation in a live, clinical environment requires strategies for organisational change management to overcome inertia, and allay concerns over lack of support and perceived threats to professional autonomy that workflow automation may bring[122] [123] [124].

3.4. Discussion

The conceptual model for the implementation of process-oriented systems comprises a distillation of the cross-cutting challenge themes that have been abstracted from 15 years of published research. It attempts to provide a concise synthesis for practitioners and implementers, by summarising the various approaches that have been proposed and implemented to date, while remaining neutral in terms of software, hardware, and knowledge/information model. It extends and generalises the model for the 'careflow' development methodology illustrated by Quaglini et al.[69] (see Figure 2.6 in Chapter 2) by showing the types of inputs and outputs required for each stage of the process and the relationship between idealised (as modelled) and actual (in enactment) clinical workflow, and how these feed back into an iterative process. The use of thematic analysis and PCA to summarise the findings of a large corpus of publications may be useful in future reviews, although further work is needed on applying and validating this technique.

This review is in many ways complementary to the recent review of workflow research by Unertl et al.[128]. In that paper, the authors identified 18 thematic categories via key phrase extraction from the papers they selected for review. A number of the themes they identified – e.g. 'Communication and collaboration', 'Idealized process for simulation', 'Design and ergonomics', 'Abstract task and process modeling', 'Temporality', 'Taxonomy

of workflow tasks' – are congruent with the themes identified in this review. In that paper, Unertl et al. also developed a conceptual framework, although the purpose of their model was to identify and relate the general concepts of workflow in the various definitions they encountered, rather than to specify a general implementation framework as in the model presented here.

In the system implementations that we reviewed, there was the assumption that real-world clinical processes are best represented by a formal model in which discrete events occur, performed by users with pre-defined roles. However, the application of computerised workflow systems to the complex, contextual nature of clinical workflow has recently been questioned[129]. Abstracting such processes into a sequence of discrete workflow steps may not capture the complex, collaborative nature of clinical processes. Some tasks may be partially, or provisionally, completed while other tasks are carried out in parallel, and new knowledge gained from downstream or parallel clinical processes may allow the remainder of the provisionally undertaken tasks to be completed or even cancelled[1]. Hardstone et al.[130], in a study of team working within a community mental health team, noted how paper-based clinical documentation supported real-world clinical workflow. Paper-based workflows supported informal working that could later be formally 'written up', and the provisional recording of information that could be finalised by the multidisciplinary team. Individual patient contact notes would be documented, but formal assessments would be done in rough, until ready to publish in authoritative, electronic form when jointly agreed as a team, as the information logged there was felt to be non-retractable.

The prevalence of paper-based clinical workflow within institutions that have implemented EHRs was also investigated by Saleem et al.[131] They found that paper provided a useful cognitive memory aid, better supported complex, longitudinal care processes and allowed more efficient multitasking. Paper allowed individual clinicians to organise their work according to their own needs and preferences, effectively allowing processes to be designed 'on the fly' (e.g. by reordering referral forms by various measure of priority, or addition of notes) — features that were not available in the EHR system. It may be, therefore, that attempts to computerise clinical workflow need to acknowledge the existence of

informal working and the use of provisional clinical decisions.

The 'semantic web' approaches to solving this 'adaptive workflow' problem (which is a concern also discussed in the general literature on workflow systems[132] [133]) have, in addition to the implementations described here, so far yielded a care pathway ontology[134] [135] which appears to share many features of older task-network models. However, the crucial distinction is that the semantic web approaches represent an 'open world' view[136] that allows new facts and relationships to be expressed without the constraint of a pre-defined schema[35], whereas earlier approaches only permit knowledge statements that are explicitly permitted by the schema. Full realisation of these approaches would require a knowledge backbone of best practice on the semantic web[134], and semi-automatic methods for transforming guideline text into a standard formalism, although recent work in this area has achieved some useful results[48][56][65].

A number of studies addressed the problem of the vagueness of concepts within text-based guidelines. The use of implicit knowledge, where terms cannot be mapped to a standard terminology or vocabulary such as SNOMED CT, either as simple, pre-coordinated concepts, or as compositional, post-coordinated concepts, presents a barrier to interoperability with EHRs. Studies that have looked at coverage of guideline terms in standardised vocabularies have found coverage of 71% (SNOMED CT, pre-operative assessment guidelines)[137]; 88% (SNOMED CT and LOINC, immunisation guidelines)[64]; and 48% (UMLS, guidelines for treatment of high blood pressure and high cholesterol). All these studies noted that the guidelines lacked explicit definitions of many terms, thus requiring manual or partial mapping to vocabulary concepts.

SNOMED CT post-coordination rules may need to become more sophisticated in order to enable more mappings to guideline concepts to be made[137]: for example, the ability to combine anatomical sites with measurements (important for recording pathological findings, such as size of colorectal mass). Also, guideline terms, even if they can be entirely mapped to a terminology, must also be mapped to terms used in the patient record. Combining the post-coordination features of SNOMED CT with the term-coverage of UMLS may improve concept mapping[137].

Even when implicit knowledge in guidelines has been made explicit and mapped to a standardised terminology, the encoded knowledge model needs to be localised to match local practice. In particular, guideline concepts need to be mapped to available fields in the local EHR. However, a one-to-one mapping may not be possible: EHR support for structured data capture of the concepts may be spread over a number of fields, or may be missing entirely[66]. This is perhaps not surprising: a one-size-fits all EHR that allows structured data capture of all potential guideline concepts would probably be unusable and agreement on the schema would probably never be agreed. As we noted in the introductory chapter (Chapter 1), around 50% of the useful information in the EHR still remains in free text fields[138]; it may be infeasible to capture it all as structured data at the point of care for process-oriented decision support purposes, although recently companies such as Clinithink[3] and Nuance[4] have been using natural language processing and speech recognition techniques in order to attempt to solve this.

With the advent of clinical information models such as HL7 RIM and CDA, and IHE interoperability protocols such as XDS, we have noted the transition from the reporting of standalone systems to the reporting of service-based, enterprise integration architectures[103]. Whether these architectures, in combination with semantic web approaches, can solve the problem of clinical workflow integration and adaptation, is an area of current research[139]. The implementation of adaptive, multi-agent, semantically aware, service-oriented workflows, incorporating formal models of clinical guidelines, appears to be a major challenge[140].

In a recently published commentary in the *Journal of Biomedical Informatics*, Sen et al.[141] also noted the transition from diverse, standalone and ad hoc CDSS systems, through to service-based architectures converging towards centralised, standardised integrated architectures. Interestingly, though, they did not note the recent trends in distributed and less centralised semantic web architectures identified in this review.

[3] http://www.clinithink.com/
[4] http://www.nuance.co.uk/for-healthcare/by-solutions/clinical-documentation/index.htm

3.4.1. Review limitations

By focusing on descriptive studies to provide a rich picture of a process, we have not considered any measures of the effect of these systems on clinical practice, nor which parts of the process are associated with successful outcomes. However, a recent systematic review of the effectiveness of clinical pathways noted that the poor quality of reporting of the implementation process prevented analysis of factors that might be critical to success[11]. In the prototype development and system architecture studies we selected, the implementation process was generally well described, but evaluations tended to be formative and weak. Future reporting of such systems should contain a richer evaluation of both the process and the outcome, to enable future systematic reviews to consider both aspects, and to determine the relative importance of the challenge themes identified.

This review has only considered studies that were published in English in peer-reviewed journals or conference proceedings published between 1995 and 2010. Consideration of information from additional sources, for example, public- and privately funded research consortia, technical reports, and professional textbooks, might lead to additional insights.

One criticism of attempting to carry out a meta-synthesis of qualitative research is that the results may have little validity, as they are based on a third level of interpretation, far removed from the original event[7]. Although development of the challenge themes was based on those identified in an earlier expert opinion paper[5], these would need to be validated by other researchers to improve the reliability and validity of the findings of this review.

3.5. Summary

This chapter has surveyed the literature on the computerisation of clinical workflow, guidelines, and pathways and the underlying, cross-cutting themes that describe the challenges to implementing process-oriented health information systems have been extracted using thematic analysis techniques. Principal component analysis has been used to cluster these themes into ten distinct groups, from which a conceptual model of the implementation process was developed.

From the development of systems supporting individual clinical decisions, Web technologies are now being used to integrate guidelines, workflows, pathways and clinical decision support towards implementation of adaptive care pathways. Such systems incorporate formal models, shared clinical knowledge resources, organisational ontologies and workflow management systems. Combining these promising architectures with more formal modelling of clinical goals and care plans, and a method for recording where deviations from these plans have occurred ('variance'), may offer the best way forward for implementation. The challenge is to provide adaptive workflow that allows dynamic modification of tasks, roles, and activity sequencing in response to changing conditions. However, the evidence-base for these process-oriented systems is in its infancy – perhaps because they are potential enablers of intervention, rather than interventions in themselves. Nevertheless, these systems need to be evaluated on a wider scale within clinical settings.

Whichever architecture is adopted, however, the core problems remain of how to model clinical processes, translating guideline and pathway text into a computer-interpretable model, and mapping the concepts contained therein to data in the EHR. To date, these problems have largely been addressed by a collaborative, time-consuming and iterative process of manually developing IF...THEN rules from guideline, protocols and pathway documents, developing clinical algorithms and flowcharts that can then be encoded in a CIG or workflow model, and creating mapping ontologies between an encoded guideline and the EHR schema.

Recent attempts to address this 'knowledge acquisition bottleneck' through the use of natural language processing and information extraction techniques have yielded promising results, and warrant further exploration. The second half of this thesis considers how to apply these techniques to concept mapping terms both in guidelines and in the free text of clinical notes in the EHR, extracting process knowledge from guidelines and clinical notes, and identifying narrative event chains in the latter. The methodology is described in the following Chapter 4, and the method is implemented in an open source, modular framework, described and evaluated in Chapters 5 to 8.

Part II.

Identifying and extracting care processes from the clinical narrative

4. Methodology for identification and extraction of clinical concepts, events and processes

4.1. Introduction

In Part I, we saw that one of the key tasks in integrating a guideline-based clinical decision support system with an electronic health record is to map clinical terms contained both in guidelines and patient notes to a common, controlled terminology. This can facilitate the provision of point-of-care recommendations and allows encoded guidelines to be shared across institutions. In Chapter 3 we identified that, despite a extensive body of research into modelling and formalising clinical processes, some core problems remain, namely how to extract the conceptual and process information that resides in the unstructured text of guideline documents, and how to map this information to concepts and process knowledge in the patient record, much of which resides in the system not as structured fields, but as free text. In this chapter, a methodology for developing a framework to extract this knowledge is described.

Extracting structured information from unstructured or semi-structured text is known as *information extraction* (IE). For some semi-structured texts, such as tabular material in HTML or CSV format, techniques such as template filling can be used, which involve patterns that exploit the structures in the data. For example, in a table, the column headings can be used to identify the type of data in each column, or text fields separated by a delimiter can be split into data types based on the position of the delimiter in the

line of text – for example, the text before the first tab might be a drug, the text before the second tab the dosage, and the text before the third tab the frequency. However, across texts at large, such templates cannot be generalised, and so *natural language processing* (NLP) techniques need to be employed.

As with the formal models of clinical and workflows processes we considered in Part I, computational processing of natural language text requires some formal representation of the text to be processed. The types of representations can be divided into 4 types, although in practice they are often combined and the distinction is not always clear-cut:

1. Lexical language models

    a) *Regular expressions and state machines*: Natural language is treated at its simplest, surface level: as a sequence of characters and strings. Patterns of characters and strings can be formalised as a finite state machine (see Chapter 2), which can be serialised in a type of algebraic notation known as a regular expression. For example, a sequence of alphabetical characters following a word boundary (a space or punctuation), optionally beginning with a single uppercase letter followed by one or more lowercase letters, then a consonant followed by the suffix '*-itis*', can be represented with the regular expression:

    ```
    \b[A-Z]?[a-z]+[^aou]itis\b
    ```

    Input text sequences are tested against the expression to see if there is a match. This expression would match '*rhinusitis*', '*Bronchitis*', '*arthritis*', but not '*coitis*'.

    b) *Word and sentence tokenisation*: Text is split into tokens that correspond to words and sentences in the language being processed. In English, this can be partially achieved by identifying white space and punctuation via regular expressions, but exceptions for dealing with abbreviations and numbers containing a decimal point need to be accounted for, either via manually created rules or using a supervised learning algorithm with a statistical model[1].

    c) *Morphological analysis*: Each token is assigned its part of speech (POS): noun,

verb, adjective, adverb etc, using either dictionaries and manually created rules (e.g. in English, there is a defined set of determiners and pronouns, many adverbs end in '*-ly*' etc), or via supervised learning algorithm. Similarly, words can be broken down into their lemmas, stems (or *morphemes*) and affixes (prefixes and suffixes) using a morphological parser, using a finite state machine operating over words[1], or a stemming algorithm such as Porter[2].

2. Syntactic language models

   a) *Context free grammars (CFG)*: grammatical text can be generated by *production rules*, and similarly, text can be decomposed into its constituent grammatical units as a hierarchical parse tree, or into linear sequences via shallow parsing (*phrase chunking*). For example, a CFG for a sentence will typically consists of noun phrases (NP), verb phrases (VP) and prepositional phrases (PP), each of which, depending on the sentence complexity, may be broken down further into smaller NP, VPs, and PPs, until the terminal leaf nodes of the parse tree consist of atomic POS units such as determiners, nouns, prepositions, verbs etc. In English, a set of complex rules for identifying the head of a phrase (the word around which the phrase is built) has been identified[1], but in practical NLP applications this is often simplified to selecting the right-most noun of a NP, PP or VP.

   b) *Dependency grammars*: dependency grammars are less concerned with parts of speech and more with the role a word plays in a sentence or clause and the typed relations between words, such as nominal subject, direct or indirect object, prepositional modifier, temporal modifier. These relations are modelled as labelled arcs between nodes of a dependency parse tree, which can be generated from a standard CFG parse tree via an algorithm[1]. Alternatively, these relations can be identified via a shallow parse from the position of words relative to the verb in each phrase chunk, either through rules or learned via a probabilistic model[3].

3. Probabilistic language models

a) *Bag-of-words*: text is modelled as a set of words in which the order of words in the text is not preserved. Words are indexed and phrases, sentences or complete documents are represented by vectors of length i, where i is the number of index entries corresponding to the number of distinct words across the texts. Each vector represents the frequency distribution of index entries in that text. For example, for the two sentences

> 'For patients who are pregnant, an ACE inhibitor should not be prescribed'
>
> 'An ACE inhibitor is prescribed only if the patient is not pregnant'

the following index would be generated (in practice '*patients*' and '*patient*' would be a single index entry after stemming each word):

```
{ACE: 1, an: 2, are: 3, be: 4, For: 5, if: 6,
inhibitor: 7, is: 8, not: 9, only: 10,
patient: 11, patients: 12, pregnant: 13,
prescribed: 14, should: 15, the: 16, who: 17}
```

resulting in the following vectors:

```
[1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1]
[1, 1, 0, 0, 0, 1, 1, 2, 1, 1, 1, 0, 1, 1, 0, 1, 0]
```

The similarity of two texts can then be scored by calculating the dot product of their vectors[1][1]. Vectors can be combined into a matrix to represent the entire corpus (collection of texts), over which calculations can be performed to identify topics in the corpus.

b) *N-grams*: Unlike bag-of-words, an n-gram is an ordered sequence of n words. Typically bigrams ($n = 2$) and trigrams ($n = 3$) are used. n-grams are collected

[1] In practice, vector entries are weighted according to some measure, e.g. inverse document frequency to give preference to words occurring frequently in a given text but less frequently in the texts as a whole.

from a corpus of documents, by counting sequences of consecutive words (normalised for cliticisation, e.g. *what's* → *what is*) using a sliding window of size n. A probability model is created based on the collocation frequency of those words, in order to predict the next word in a sequence of $n-1$ previous words, using the Markov assumption that the probability of word w being the next word depends only on the previous word: $p(w_n|w_{n-1}, w_{n-2}...) = p(w_n|w_{n-1})$.

4. Conceptual and semantic models

   a) *Word senses*: words are classified according to their meaning. A word paired with its specific meaning sense is known as a *lexeme*[1]. For example, 'stroke' might refer to the verb 'to stroke', as a *synonym* for 'to caress', or be used as a *metonym* (figure of speech or alias) for 'cerebrovascular accident', depending on context. Lexical and semantic relations between English words in a variety of domains have been classified in a database called *WordNet*[4], which is widely used in NLP tasks (see Chapter 8), and can assist in *word sense disambiguation* (WSD), which is a key problem in clinical natural language processing (for example, expansion of abbreviations; see Chapter 7).

   b) *Semantic role labelling (SRL)*: Verbs are classified as *events* and their participant noun phrases as event *participants*, each with a thematic role[1]. Thematics roles such as `AGENT` (causer of an event), `THEME` (thing affected by the event) and `INSTRUMENT` (object used by the `AGENT` in the event) often correspond to subject, object and indirect object dependency relations. Semantic roles pertinent to clinical NLP include `PATIENT` or `EXPERIENCER` (often analogous to `THEME` but implies some state change), `TIME` (the time that the event occurred), `INSTRUMENT` and `GOAL`. The FrameNet project[5], a database of over 170,000 role-labelled sentences, aims to model hierarchical structures of related events and participants to facilitate the development of SRL systems.

4.2. Research methodology

In this research, we combine the use of public-domain ontologies (see Chapters 2 and 3) with lexical, shallow syntactic (via phrase chunking: Chapter 5, and some dependency relations: Chapter 8) and semantic models with hand-crafted patterns, developed iteratively through analysis of linguistic patterns that occur in ontology terms and across clinical texts in general. The exception is the work described in Chapter 8, where patterns are developed from a subset of an existing corpus (the training set) and then tested on a previously unseen subset from the same corpus (the test set; see Section 4.2.2 below). Overall, the approach used throughout this work is to apply knowledge-based principles to identify general patterns that occur in the domain, and then evaluate them for specific task performance (term identification and mapping, negation and possibility, abbreviation expansion, spelling correction, coreference resolution) against a number of corpora.

4.2.1. Development methodology

Clinical NLP research over the last 15 years or so has been dominated by statistical and machine learning approaches[6]. Recent advances in the application of supervised machine learning techniques to the biomedical and clinical domains – particularly conditional random fields (CRF) and support vector machines (SVM) – have led to the development of effective systems for concept and relation identification, although these require significant feature engineering effort (see overview in Uzuner et al.[7]). The focus of the method in this work, however, is on iterative development of explicit, human- and machine-readable patterns and rules for the creation of general-purpose, interoperable components. The justifications for this approach include:

- Features that lead to useful results, and that have been identified via experimenting with rules and patterns, can still be used as input to a machine-learning process in future work. Rule-based and machine learning approaches co-exist in many hybrid systems: the combination of lexical rules with external knowledge resources has value in providing input to, or post-processing the output of, statistical and machine-learning approaches, particularly when available training data is sparse[7].

For example, Patrick et al.[8] used dictionary lookup to expand abbreviations and add UMLS classes and identifiers as token features, which were used as input to a CRF classifier that augmented a rule-based system.

- Supervised learning techniques typically require a large amount of hand-labelled training data. Although such data is available for participating in specific biomedical or clinical NLP challenges (such as the annual BioNLP and i2b2 Shared Challenges – corpora from which are used for evaluating this work), the types of labels and features used in these corpora are specific to the challenge. It is not clear how a classifier trained on such data would generalise to other data labelled differently without requiring retraining on the new data. The goal of this work, however, it to develop patterns and components that can be used to solve a number of problems across a range of corpora, with flexibility in the types of concepts and relations identified.

- Some of the best-performing systems for solving specific NLP problems in the general domain, such as identification and normalisation of temporal expressions (see Chapter 5) and resolution of coreference (see Chapter 8) are entirely rule-based.

- Earlier rule-based systems fell out of fashion as, over time, they resulted in monolithic, unmanageable sets of handwritten rules that could lead to non-deterministic behaviour[6]. However, advances in Web-based, distributed and modular software architectures (see Chapter 3), dynamically loaded libraries and 'pipeline' based NLP architectures (see Chapter 5) has made the development of self-contained but extensible rule-sets more feasible, potentially avoiding the problem of earlier systems where many interdependent rules became unmaintainable. Clearly defined rule-sets can be extended and modified by others to suit local needs; and rules can be verified for consistency and conflict, and validated against organisational requirements.

The use of lexical and syntactic patterns and morphemes for processing clinical text is not a new idea. Back in 1969, Pratt and Pacak[9] described some principles for mapping text to terms in the Systematized Nomenclature of Pathology (a forerunner to SNOMED),

noting that clinical terminology is highly compositional, consisting of Latin, Greek and English morphemes. They described morphological and lexical transformation rules for identifying semantically analogous phrases (e.g. '*atrophy of muscle*' → '*muscular atrophy*'). The multi-decade Linguistic String Project[10], which led to the commercial MedLEE system, mapped clinical text to SNOMED concepts by identifying word sequences from a dictionary of words mapped to lexical categories such as `PART`, `AREA`, `INDICATION`, `AMOUNT`.

More recently, Patrick et al.[11] used regular expressions, the UMLS Specialist Lexicon, and an n-gram model to identify and map terms occurring in the free text of EHR patient notes to SNOMED CT. Kaiser et al.[12] used the MetaMap Transfer application (MMTx – see Chapter 5), the UMLS Semantic Network and syntactic patterns to identify clinical actions and conditional statements in guidelines (see also Chapter 3). Serban et al.[13] identified a number of linguistic patterns useful for the creation of templates to extract terms and processes from the text of clinical guidelines. For example statements such as:

> In the event of [pregnancy]$_{\text{med\_context}}$, [patients with diabetes]$_{\text{target\_group}}$ [should]$_{\text{recommendation\_op}}$ be [prescribed calcium channel blocker]$_{\text{med\_action}}$

and

> For [diabetic patients]$_{\text{target\_group}}$ with [kidney damage]$_{\text{med\_context}}$ the [blood pressure target is 130/80]$_{\text{med\_goal}}$

can be generalised by respective patterns

```
(med_context, target_group, recommendation_operator, med_action)
(target_group, med_context, med_goal)
```

The methodology used in the current work recapitulates and extends some of the ideas described by Pratt, Sager, Patrick, Kaiser, Serban and others, and implements and evaluates them in a systematic way. In order to develop tools and techniques that can be implemented in a distributed, web-based environment (see Chapter 3), it makes sense to use an existing framework that supports this implementation. In this work, the open-source General Architecture for Text Engineering (GATE)[14] framework will be used for the natural language processing component. GATE was chosen for the following reasons:

- its pipeline-based approach allows self-contained components that individually may use rule, dictionary, or statistical and machine learning approaches, to be integrated and reconfigured for different tasks;
- it is well supported, with a large, international community of developers and users;
- it is open-source software with a well-documented application programming interface (API) which facilitates customisation and extension of the core software.

Identification of processes of care in unstructured clinical texts involves the following tasks[15][16]:

- identification and post-coordination of clinical terms;
- identification of term aliases and abbreviations that point to the same, previously described real-world entity (coreference);
- identification of events and action phrases (e.g. '*perform*', '*prescribe*');
- identification of conditional phrases and clinical rules ('*if ... then*');
- identification of dose unit terms (e.g '*3 mg*');
- identification of temporal expressions (e.g. '*24 h*', '*two days*') and temporal constraints (e.g. '*for at least 2 weeks*', '*no more than 3 days*');
- identification of negation and possibility ('*absence of symptoms*', '*pain not controlled by medication*', '*surgery may be appropriate*').

In this work, we consider each of these tasks. Chapter 5 deals with all except the second task, which is dealt with exclusively by Chapters 7 and 8, and Chapter 6 considers an alternative approach to the first task.

4.2.2. Evaluation methodology

The performance of the tools and techniques, developed during this research to identify concepts and processes, are evaluated against a range of texts in the clinical and biomedical

domain: clinical guideline documents, MedLine abstracts, discharge summaries, progress notes, surgical, pathology and radiology reports. Where available, performance is evaluated against manually annotated corpora. As noted in Chapter 1, the purpose of creating these labelled corpora is to facilitate the development of systems that can recognise such concepts and relations automatically, in order to identify opportunities for decision support.

In creation of these corpora, a team of domain experts are typically given the task of labelling concepts and relations in accordance with a guidance document produced by a research team. For example, the task might be to identify all phrases that contain human anatomical terms, or to identify mentions of clinical procedures, disease concepts, or medications. In addition to labelling the concepts, domain experts – *annotators* — may be asked to assign attributes (*features*) to each term, such as whether it is expressed in a positive, negative or possible context (e.g. the non-appearance of a possible symptom, the rejection of a finding, the possibility of a diagnosis), or whether the term relates to a described process that has occurred, is planned to occur, or may occur. Agreement between different annotators will vary; typically, the publishers of the corpora will cite a figure for the inter-annotator agreement (IAA)[17] using Cohen's kappa (κ)[18] for agreement between pairs of annotators or Fleiss' kappa[19] for more than 2 annotators, where

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the observed proportion of cases where all annotators agree on a classification and p_e is the proportion of agreement expected by chance. In some cases (e.g. Appendix 2 of [20]), a simplified measure is used that calculates a score based on the ratio of cases where there is agreement to case where there is disagreement:

$$IAA = \frac{2 \times N_{match}}{2 \times N_{match} + N_{nonmatch}}$$

where N_{match} is the number of cases in which all annotators agree and $N_{nonmatch}$ is the number of cases in which they disagree. The aim of an automated system would be to at

least match the IAA score, i.e. perform at least as well as domain experts agree. In practice, the corpus will be adjudicated – another set of domain experts will curate the corpus and resolve cases where there is disagreement between annotators, producing a single set of documents that represents the corrected version of all the individual annotators' work. Such a corpus is referred to as a *gold standard*, and the aim of an automated system would be to match this gold standard as closely as possible.

Where labelled corpora are not available, we make use of '*silver standard*' data by creating an annotated corpus using an automated tool that has been accepted by the research community as producing high-quality output and against which other approaches should be measured. In this case, the National Library of Medicine's concept recogniser, MetaMap[21], is used to create the silver standard, as it is considered to be the reference tool for concept identification in the biomedical domain[22].

The evaluation measures used in this research will be the standard measures used in the IE and NLP fields: *precision*, *recall*, and F_1-measure. Measurement of recall (also referred to as sensitivity) determines the ratio of correctly identified annotations ('true positives') to the total number of annotations in the gold standard version (i.e the sum of 'true positives' and 'false negatives'; see Equation 4.1). Measurement of precision (also referred to as positive predictive value) determines the correctness of the annotations identified; it is the ratio of true positives to all annotations (correctly and incorrectly) identified (Equation 4.2). F_1-measure is a measure of the balanced performance of the system, defined as the harmonic mean of recall and precision (Equation 4.3)[2].

$$recall = \frac{TP}{TP + FN} \tag{4.1}$$

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

where TP is the number of true positives (correct matches), FN is the number of false negatives (missed matches), FP is the number of false positives (incorrect matches).

[2]The '1' refers to the value of the weighting coefficient β between precision and recall; more generally, $F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall}$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.3}$$

In practice, precision, recall and F_1-measure can be calculated as either strict, lenient, partial or average scores. In strict scoring, in addition to matching the classification of a concept (i.e. a concept classified as `Procedure` by the system but as `Test` in the gold standard would not score as a true positive), the extent of a system-generated annotation must exactly match that of the corresponding gold standard annotation. For example, if the system identifies an `AnatomicalTerm` concept of '*femur*' starting at character position 10 from the beginning of the document (the *start offset*) and ending at character position 15 (the *end offset*), but the gold standard has an `AnatomicalTerm` concept of '*the femur*' between offsets 6 and 15, this would not be scored as a true positive and would be counted as a false positive. With lenient scoring, however, system-generated extents that overlap with the corresponding concept in the gold standard do count as true positives. With partial scoring, matching concepts with exactly matching extents are each scored 1 and overlapping extents are each scored 0.5. With average scoring, the mean of strict and lenient scores are calculated. Unless otherwise specified, strict scoring is used in evaluating this work.

However, as they stand, Equations 4.1 to 4.3 only calculate scores for matching a single class of annotation or feature in a single document. In practice, we match multiple classes (e.g. `Procedure`, `Disease`, `Treatment`, `Test` concepts) across multiple documents. Across a corpus and across classes, precision, recall and F_1-measure can be calculated as either a micro-average or a macro-average. Micro-averaging sums individual numerators and divides by the sums of individual denominators (Equations 4.4–4.6):

$$precision_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + FP_i} \tag{4.4}$$

$$recall_{micro} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + TP_i} \tag{4.5}$$

$$F_1^{micro} = 2 \times \frac{precision_{micro} \times recall_{micro}}{precision_{micro} + recall_{micro}} \tag{4.6}$$

where n is the number of classes being evaluated, or, for evaluating a single class over a number of documents, the number of documents in the corpus. Whereas macro-average sums individual scores and divides by the number of scores (Equations 4.7–4.9):

$$precision_{macro} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i} \tag{4.7}$$

$$recall_{macro} = \frac{1}{n} \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{4.8}$$

$$F_1^{macro} = 2 \times \frac{precision_{macro} \times recall_{macro}}{precision_{macro} + recall_{macro}} \tag{4.9}$$

Micro-averaging over classes gives performance measures over all classes for the corpus as a whole, whereas macro-averaging over classes would give a measure of performance for a typical annotation class. (Similarly, micro-average scores for a single class over each document gives average performance measures for that class in a document, and macro-averaging for a single class over the corpus gives average performance measures for that class over the corpus; this makes more sense when considering the scores for annotation features, i.e. given a matched annotation, what is the likelihood that the features are correct – see Chapter 5.) In this work, unless otherwise specified, micro-average measures are reported.

In their model for evaluation of information system success, DeLone and McLean[23] (D&M) consider six inter-related axes for evaluation:

1. *System quality*: ease of use, functionality, reliability, flexibility, and portability
2. *Information quality*: accuracy, timeliness, completeness, relevance, and consistency
3. *System use*: frequency of use, number of times system accessed, the degree to which users depend on the system

4. *User satisfaction*: users' perceptions of how important and useful the system is for achieving their work goals

5. *Individual impact*: quality of decision making, job performance, task productivity

6. *Organisation impact*: time and cost savings, improved customer outcomes

Within this model, the evaluation of the current work fits within the second axis of information quality, in that the techniques and tools developed here transform unstructured text into semantically enriched data from which decision support information may be more easily extracted. In terms of D&M, we measure the accuracy, completeness and consistency of the data generated by the components developed. In addition to the measures discussed above, we also evaluate the performance of the components developed in terms of their processing speed (timeliness within the D&M model) on the evaluation corpora (see Chapter 5). Furthermore, as the components developed in this work have been released as open source software, formative evaluation of system use might be made in terms of numbers of downloads or the number of other research or commercial projects in which these components are subsequently used. This is discussed in Chapter 9.

4.2.3. Data collection

City University and the British Medical Journal (BMJ) have signed a collaboration and Non-disclosure Agreement (NDA) to allow use of their care pathway and clinical guideline data from their Best Practice and Clinical Evidence products. A full data set for both products has been provided by the BMJ. Additional clinical guideline data will be harvested from National Institute for Clinical Excellence (NICE) guidelines, which can be reproduced for educational and not-for-profit uses, and from the US National Guideline Clearing House.

Deidentified clinical records have been provided by the i2b2 National Center for Biomedical Computing funded in part by grant numbers U54LM008748 and 2U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data supported by the VA Salt Lake City Health Care System with funding support from the

Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204 and the National Institutes of Health, National Library of Medicine under grant number R13LM010743-01. A Data Use Agreement has been signed by both parties.

4.2.4. Ethical approval

The City University Research Ethics Committee have confirmed that ethical approval is not required for this project as it involves the use of existing, anonymised data sets for which research approval has been granted by the data owners.

4.3. Summary

This chapter has outlined a development and evaluation methodology for a clinical information extraction framework. Each of the following Chapters 5 to 8 provides details of the specific methods used for addressing the domain problems tackled by that chapter. Chapter 5 describes methods and formative evaluation of the core framework modules of term identification, concept mapping, negation and possibility, temporal concepts, processes and events. Chapter 6 evaluates a method for more efficient ontology-based term identification. Chapter 7 describes a method and an evaluation of dynamically generated regular expressions for expanding and coreferencing biomedical and clinical abbreviations. Chapter 8 integrates all these methods to trace processes of care as linked chains of events through coreference resolution.

5. Development of an open-source, modular framework for clinical concept and process extraction

5.1. Introduction[1]

As we saw in Chapter 4, extensive research effort has been invested in developing and applying lexico-syntactic, knowledge-based and statistical machine learning methods to the problem of identifying clinical concepts, events, relations and process knowledge in unstructured text. Despite this effort, while there are a number of individual, open-source tools and applications for working with text in the biomedical domain (for example, for identifying genes, proteins, their associations, expressions and interactions), the number of freely available, general-purpose tools for working with clinical texts are fewer. In this chapter, we describe the development of, and provide formative evaluation for, a number of novel, interoperable components for extracting clinical concepts and process information from clinical guidelines and patient notes.

Existing tools for extracting information from biomedical texts include the Stanford Biomedical Event Parser[2], the GENIA tagger[3], ABNER[4], AbGene[5], the Penn Biotagger [6] [7], and MutationFinder[8]. MetaMap[9], from the US National Library of Medicine (NLM), which is probably the most widely used and is considered to be the reference standard tool for mapping unstructured text to concepts in the UMLS, was originally developed for retrieving and processing MEDLINE abstracts. Although its use in

[1]Some material in this paper has been published as 'A tool for enhancing MetaMap performance when annotating clinical guideline documents with UMLS concepts'[1].

its more recent incarnations has been extended to clinical texts, MetaMap only handles unformatted ASCII text as input, can handle a maximum input size of 3000 characters, and processing complex phrases can require many hours of computation[10].

Despite the development of these tools, the lack of interoperability between them has been recognised as a major barrier for researchers in biomedical and clinical natural language processing and information extraction[11][12]. Standalone tools typically require 'glue code' to work together, as they each may be written for a specific task, and may have varying input and output formats[12].

The requirement for glue code can be minimised by wrapping the tools around a standard application programming interface (API) or application framework. In the open-source arena, there are a number of such frameworks for different programming environments. Three of the most popular include the Natural Language Processing Toolkit (NLTK)[13] for the Python language, and the General Architecture for Text Engineering (GATE)[14] and the Unstructured Information Management Architecture (UIMA)[15] frameworks for the Java language. These open-source frameworks allow individual components to be used together in a processing workflow or *pipeline* in which the output of one component can be used as the input to a later component[16].

As a result, some of these individual tools have been integrated into these frameworks. The GENIA tagger, ABNER, AbGene, Penn Biotagger, and MutationFinder have been made available as GATE components[17]. Researchers at the National Centre for Text Mining (NaCTeM) have also integrated a number of these tools as components in a UIMA-based open-source system called U-Compare[11].

In the clinical domain, a well-known, open-source text processing framework is the *Mayo clinical Text Analysis and Knowledge Extraction System* (cTAKES)[18], which is also based on UIMA, integrates the NLM's Lexical Variant Generation (LVG) tools[19] (see Chapter 6), and comprises a number of components standard in the general NLP domain but trained specifically for clinical texts: sentence chunker, tokenizer, and part-of-speech (POS) tagger; and specific components for extracting clinical concepts and their coreference relations[20] (see Chapter 8). However, cTAKES has a complex installation

and configuration process[2] and does not appear to be easily used or customised by end users without programming ability.

One benefit of UIMA is that there is an integration component for MetaMap. However, the MetaMap integration with UIMA does not provide 'out of the box' clinical concept mapping for end users without configuration, knowledge of the Java programming language and the UIMA API[3].

Another open-source clinical text processing system is the *cancer Text Information Extraction System* (caTIES)[21], originally built around the GATE framework. Its focus is on concept mapping surgical-pathology reports, rather than providing a general-purpose clinical knowledge extraction pipeline. Although it is integrated with MetaMap via the MetaMap Technology Transfer (MMTx) application, MMTx has now been deprecated by the NLM and is no longer supported since the public release in 2008 of the MetaMap server software[4].

The Topaz[5] system is also based around GATE. Its main purpose is to identify respiratory conditions in free text, but makes use of modules that implement the well-known NegEx algorithm[22] for detecting negated clinical findings and ConText[23] for identifying their temporal, hypothetical and experiencer (i.e. patient or family member) contexts. NegEx uses a list of 270 'trigger terms' that may appear before or after a clinical concept (usually a disease, finding or symptom) that indicate whether the concept is possible (e.g. '*may not be ruled out*') or negative ('*no evidence of*', '*was ruled out*'), and has a reported precision and recall of 84.5% and 77.8% against a test set of 1235 concepts in 1000 sentences extracted from discharge summaries.

The Health Information Text Extraction (HITEx) system[24] is another GATE-based suite of tools, consisting of components for identifying smoking status, principal diagnosis, and discharge medication. However, it is unclear whether this project is still being actively maintained, as it was developed for an old version of the GATE framework (version 3.1) which has been significantly updated since then (currently GATE is at version 7.1).

[2] `https://wiki.nci.nih.gov/display/VKC/cTAKES+2.0+User+Install+Instructions`
[3] `http://metamap.nlm.nih.gov/README_uima.html`
[4] `http://mmtx.nlm.nih.gov/MMTx/`
[5] `http://www.dbmi.pitt.edu/blulab/resources.asp#Topaz`

Notably, the tools described above have been developed for quite specific tasks in the clinical domain. There still seems to be a need for loosely coupled, general-purpose components that can be used within an open-source framework for clinical text processing. These interoperable modules need to be able to work independently, or as part of a larger pipeline, without the need for additional software development (the 'glue code') requiring programming expertise by the end user. In the remainder of this chapter, we describe and evaluate a number of software components that aim to address these requirements.

The use of the pipeline paradigm for performing a succession of transformations on clinical texts was probably first described by Meystre and Haug[25]. In that work, pipeline steps included text segmentation, clinical problem identification, negation detection, and post-processing for local code mapping and XML generation. In this research, we build on the pipeline concept described by Meystre, but with the goal of flexibility in component sequencing – i.e. one or more components (except for any essential, generic pre-processing steps, as described below) may be omitted or even used repeatedly without 'breaking' the pipeline, and there is minimal coupling between components.

In this work, each of these components are developed for the GATE framework, as GATE provides a text processing environment that can be used 'out of the box' without requiring the end user to have programming expertise. Moreover, unlike UIMA, GATE (prior to this current work) lacked integration with MetaMap, and had no specific components for identifying basic concepts pertinent to clinical texts, such as

- quantitative concepts: number, measurement, units;
- temporal concepts: time, date, duration, age, frequency;
- process concepts: actions, events, temporal relations;
- negation and possibility;
- spelling correction;
- abbreviations.

In this chapter, we consider MetaMap integration into the GATE framework, and approaches to improve its performance. It was noted earlier that MetaMap requires input text to be segmented into management chunks (< 3000 characters) in order to be able to process it. Here, we consider different text segmentation approaches to maximise the capture of contextual information while retaining relative accuracy of annotation in comparison to default MetaMap behaviour. In addition, we describe lexico-syntactic patterns for the identification of the above quantitative, temporal and process knowledge concepts that provide the context for the clinical terms identified by MetaMap, and we evaluate the performance of the event, negation and temporal concept identification components against a corpus of clinical discharge summaries. The latter two items in the above list (spelling correction and abbreviations) are dealt with in Chapters 6 and 7.

5.2. Methods

5.2.1. Overview of the GATE framework

GATE provides native support for processing source documents in variety of text encodings and formats, including PDF, RTF, and XML, and has a mature Java API for integrating other applications – particularly those with their own Java API, such as MetaMap[6] – as components in its processing pipeline. GATE can also serialise processed text in a generic XML format, where each semantic annotation added by pipeline components is serialised as an XML element, with annotation features serialised as attributes on the corresponding element. The output XML can be transformed via an external process (e.g. an XSLT transformation) into the specific format required by the application that invokes the GATE pipeline (e.g. RDF, OWL, CSV etc).

GATE comes with a number of general-purpose components for shallow parsing of English text as part of its ANNIE[14] (A Nearly New Information Extraction system)[7]: a Tokenizer, Sentence Splitter, Part of Speech (POS) Tagger, a Morphological Analyser (for lemmatisation and identification of verb infinitives), Noun Phrase and Verb Group chun-

[6] `http://metamap.nlm.nih.gov/javaapi/javadoc`

[7] It also has components for working with most European language and Chinese, however, in this work, we only consider medical English and its neoclassical morphemes

kers, a trie-based Gazetteer component (see below), and the Java Annotation Pattern Engine (JAPE) rules engine for writing lexico-syntactic patterns over existing annotations using regular expressions. The typical ordering of these components in a pipeline is shown in Figure 5.1.

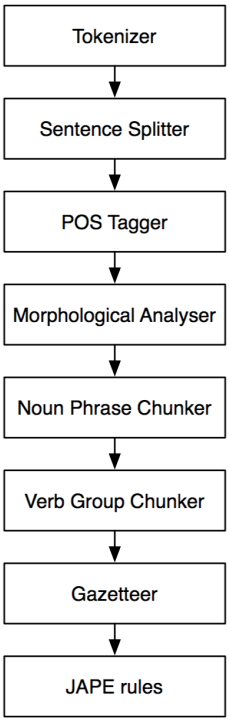


Figure 5.1.: A generic information extraction pipeline using GATE's ANNIE components

A Gazetteer comprises one or more plain text files (e.g. `anatomy1.lst`) that function as lookup lists, each of which is described in an index file that classifies each list according to major and minor types (e.g. `anatomy1.lst:human_anatomy:location`) (see Figure 5.2). The lists themselves comprise one entry per line, where each entry is a term to be looked up in the document, and can be further classified with one or more feature attributes that will be added to the annotation created when a lookup term is found in the document (see Figure 5.3).

A JAPE rules file (see Figure 5.4) consists of one or more pattern-matching rules that will 'fire' when the conditions on the left-hand side of the rule are met; the right-hand side of the rule determines the action that will be taken on firing: typically, the output of a new annotation, addition of features to an existing annotation, or other, more complex

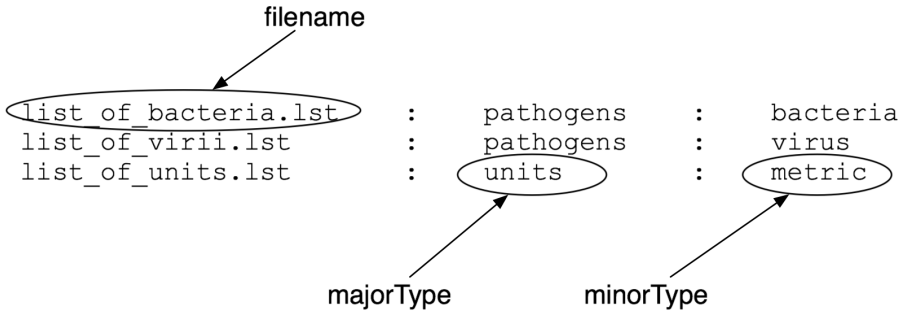


Figure 5.2.: Structure of a GATE Gazetteer list definition file (list of lists)

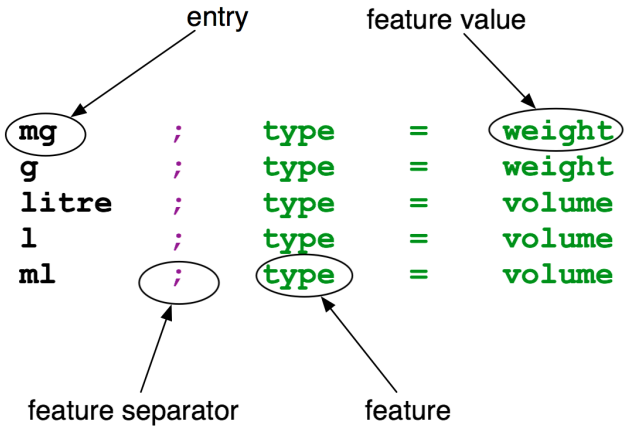


Figure 5.3.: Structure of a GATE Gazetteer list file

behaviour that can be written in Java code.

According to the review by Krauthammer & Nenadic[26], an system for identification of clinical terms typically needs to perform three tasks:

1. *recognise* the text string as a possible term (candidate term selection)
2. *classify* the candidate term (e.g. body part, disease, physiological function)
3. *map* the term to a single concept (pre-coordination) or to qualified, multiple concepts (post-coordination) within a standardised vocabulary or ontology.

The approach used for the first of these tasks is described in the following Section 5.2.2. Section 5.2.3 describes the method for tackling the second and third tasks. Sections 5.2.4 and 5.2.6 describe the method for identifying process and temporal concepts and relations.

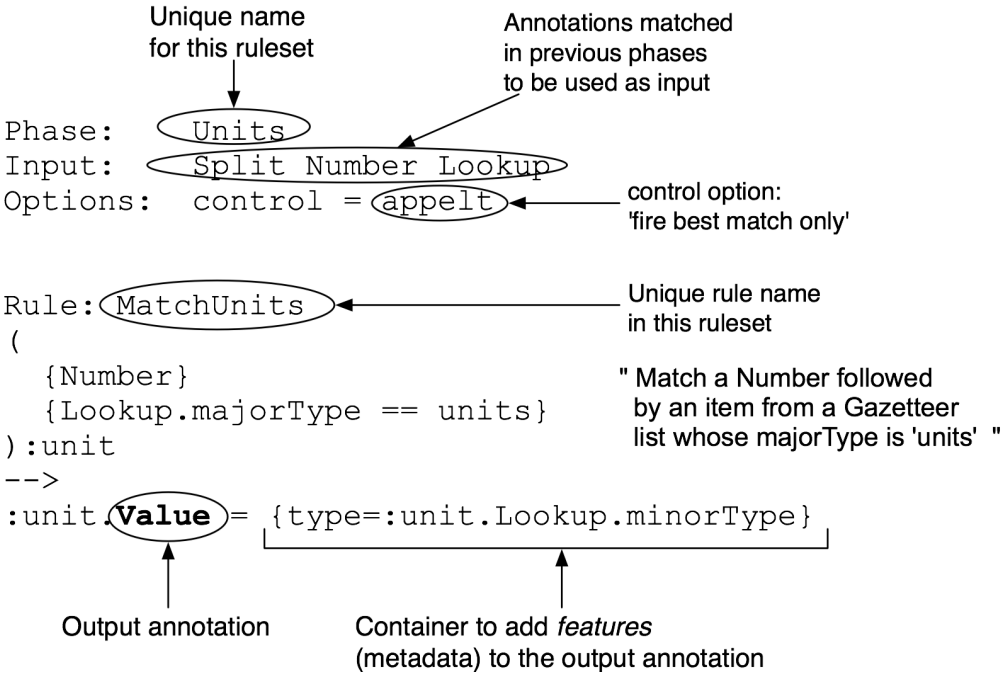


Figure 5.4.: Structure of a JAPE rules file

5.2.2. Identifying candidate clinical term phrases: text segmentation

As pointed out by Bashyam[27], simply chunking text into noun phrases (NP), verb phrases (VP) and prepositional phrases (PP) is insufficient for identifying term phrases in clinical narratives. These may consist of partial or ungrammatical phrases, or consist of semantic units comprising more than one syntactic phrase type. In clinical texts, noun modifiers may often be verb past participles, such as '*increased blood pressure*' or '*impaired glucose tolerance*', which a typical phrase chunker would identify as VPs and split them into a verb group (VG) ('*increased*', '*impaired*') and a NP ('*blood pressure*', '*glucose tolerance*'), whereas we wish to identify them as a single lexical unit for the purposes of candidate term identification. Similarly, clinical concepts may be appear as an adjective or gerund acting as a noun ('*the elderly*', '*sneezing*') or as a noun followed by an infinitive postmodifier ('*refusal to eat*', '*failure to thrive*').

To address this, the following approach to text segmentation for phrase identification was taken:

- Start with the output of a standard NP and VG chunker (in this case, the ANNIE

components described above).

- Write JAPE rules to modify the output of these chunkers to identify phrases that has a verb modifier with an '-ed' or '-ing' affix.
- Identify adjectives and gerunds not attached to a NP as a potential term if preceded by a determiner and not followed by a noun.
- Identify '`is_a`' phrases where the object is an adjective or a verb with an '*-ed*' or '*-ing*' affix and turn this into an equivalent NP. For example, '*treatment was effective*' → '*effective treatment*', '*patient is disabled*' → '*disabled patient*', '*no bleeding was evident*' → '*no evident bleeding*'.
- Expand adjectival phrases joined to a noun phrase with a coordinating conjunction. For example '*mild, moderate and severe hypertension* → '*mild hypertension*, *moderate hypertension* and *severe hypertension*'.
- Identify prepositional phrases comprising phrases identified by the above steps joined by a positional preposition (*of*, *on*, *in*, *under*, etc). Segmenting the text into both NPs and PPs allows contextual information to be captured, for example '*pain on the right side of the chest*', '*family history of congestive heart failure*', '*no evidence of cardiovascular disease*', '*absence of pulse*'.

5.2.3. Identifying and mapping clinical concepts: MetaMap integration and performance improvements

Using the MetaMap Java API, a software component was developed that integrates MetaMap with GATE, in which, by default, text is normalised and chunked into blank-line delimited segments. These text segments are submitted to MetaMap server (or multiple servers for parallel processing) and the results converted to GATE annotations and features for further processing.

Terms containing diacritics are stored in normalised ASCII form within the UMLS source vocabularies (e.g. Angstrom, Guillain-Barre etc). As MetaMap can only process ASCII data, it returns term positions as byte offsets relative to the start of the phrase.

Annotation offsets in GATE are given as encoding-dependent character offsets. Thus, multi-byte UTF-8 data in the source document payload needs to be normalised to its single-byte, ASCII equivalent so that MetaMap can map it to the correct term in UMLS (e.g. Ångström → Angstrom, Guillain-Barré → Guillain-Barre).

To accommodate this, UTF-8 and ISO- 8859-1 data in the payload is normalized to its ASCII equivalent by using the `java.text.Normalizer` class[8] which implements the standardised Unicode Normalization Forms[28] to 1) convert the string to its Roman equivalent plus diacritic, 2) strip the diacritic, and 3) convert the resulting string to an ASCII byte stream from which to create a new ASCII-encoded string that forms the input to MetaMap. This process allows terms containing diacritics to be correctly recognised and mapped by MetaMap.

The plugin provides a number of features designed to optimise the processing of large documents. Although by default the document payload that forms the input to MetaMap is chunked into line-break delimited segments, this can be modified to one of the following:

- the content of user-defined input annotations (identified by an upstream process, such as the candidate term phrases described in Section 5.2.2), or specified elements in the original XML markup;
- only distinct instances of the string content of each input annotation, with remaining instances linked back to the first mapped term (i.e. coreferencing – see Chapter 8);
- user-defined features on each input annotation, rather than the original source data. For example, an upstream process might normalise prepositional phrases (see Section 5.2.2 such as '*cancer of the lung*' → '*lung cancer*', or verb phrases such as '*pain is severe*' → '*severe pain*', and store the result as a feature on the term. This allows these grammatical variants to be treated as equivalent for the purposes of coreferencing;
- content stripped of leading determiners and possessives: for example '*He fractured his right arm*' → '*fractured right arm*';

[8] `http://java.sun.com/javase/6/docs/api/java/text/Normalizer.html`

- input annotations that do not occur within user-defined section: for example, if only the Recommendations section of a guideline is to be processed, this can be specified here;
- input annotations that do not contain user-defined elements. For example, if we wish to ignore input phrase chunks that contain or are coincident with a temporal of process concept (see Section 5.2.6 below), then these can be specified here.

5.2.4. Identifying process concepts, events and relations

As we saw in Chapter 4, previous researchers have identified specific verb groups for identifying process concepts in clinical guidelines, such as *activate*, *perform*, *prescribe*, *treat*, either from hand-created lists of verbs extracted from guideline documents[29], or from the UMLS Semantic Network[30] and an online thesaurus[31]. Here, we consider general verb group patterns that identify processes ruled in or ruled out, or actions performed, or to be performed, as part of the care process – for example, parsing out goals and rules from policy statements or guidelines (see Chapter 2, Section 2.3.4) – but we defer identification of specific verbs within these patterns required for specific tasks (see Chapter 8).

TimeML considers an *event* to be some situation that occurs in time, and defines events as 'tensed or untensed verbs, nominalizations, adjectives, predicative clauses, or prepositional phrases'[32]. In the case of clinical texts, events can be considered to be clinical concepts that describe a patient state, such as a symptom, disorder or disease (typically grouped as `Problem` in the evaluation corpora used in this work), a procedure, such as a test or treatment, or some process or occurrence that affects the patient, such as admission, discharge, referral etc, and their corresponding verb forms. Therefore in addition to problem, test and treatment-type concepts, we also consider verb groups to be potential events.

The ANNIE Verb Group (VG) chunker identifies sequences of verbs, including negations and modals. For example the phrase '*may not be appropriate to prescribe*' will be identified as two VGs: '*may not be*' with negated ('not'), modal ('may') features and '*to prescribe*' with an infinitive tense feature. Similarly, consider the phrases:

[Diazepam]$_{\text{NP}}$ [may not be]$_{\text{VG.be,modal,neg}}$ [appropriate]$_{\text{JJ}}$ [to prescribe]$_{\text{VG.inf}}$ [to]$_{\text{IN}}$ [this patient group]$_{\text{NP}}$

[Infection]$_{\text{NP}}$ [seems]$_{\text{VG.seem,modal}}$ [unlikely]$_{\text{JJ}}$ [to have been]$_{\text{VG.be,perf}}$ [the cause of these symptoms]$_{\text{PP}}$

[These symptoms]$_{\text{NP}}$ [may not have been]$_{\text{VG.be,modal,neg}}$ [related]$_{\text{VG}}$ [to]$_{\text{IN}}$ [the earlier infection]$_{\text{NP}}$

[The incision]$_{\text{NP}}$ [was made]$_{\text{VG}}$ [just]$_{\text{RB}}$ [inferior]$_{\text{JJ}}$ [to]$_{\text{IN}}$ [the edge of the tumor]$_{\text{PP}}$.

where JJ = adjective, RB = adverb, PP = prepositional phrase, NP = noun phrase, IN = preposition, inf = infinitive tense, perf = perfect tense. The relations between the above NPs/PPs can be generalised by the following patterns:

```
NP|PP VG.modal JJ VG.inf IN NP|PP
NP|PP VG.modal JJ VG.perf NP|PP
NP|PP VG.modal VG IN NP|PP
NP|PP VG RB JJ IN NP|PP
```

These subject-predicate-object patterns suggest a more general lexical pattern for identifying VG predicates that specify some clinical action or process, either performed, postulated, or planned:

```
(VG.modal | VG){1, 2} RB? JJ* (VG.inf | VG.perf | IN){1,2}
```

where |, ?, *, $\{n,m\}$ represent standard regular expression Kleene operations for specifying occurrence.

Lists of words and phrases expressing temporal relations between terms were created and mapped, where possible, to their TimeML `TLINK` relation types of SIMULTANEOUS/-DURING/OVERLAP; BEFORE/BEFORE_OVERLAP/ENDED_BY; and AFTER/BEGUN_BY. For example, the compound preposition '*prior to*' signifies a BEFORE relation between two events, such as 'patient complained of [chest pains]$_{\text{Problem}}$ *prior to* [admission]$_{\text{Event}}$'

5.2.5. Identifying negation and possibility of concepts and events

MetaMap includes an implementation of the NegEx algorithm[22]. In the current work, the text segmentation approach described above allows the words that may indicate negation or possibility surrounding the target noun phrase to be captured and thus processed by NegEx within MetaMap. For example, prepositional phrase chunking will capture '*no evidence of ...*', '*absence of ...*', verb group chunking will capture '*... may be excluded*'. However, if the MetaMap integration component is not used (see Chapter 6), a separate component for identifying negated or possible findings is required. While NegEx is available as a separate GATE framework component, it was developed for an earlier version (4.x), it requires a number of sub-components to be installed and instantiated in a specific order in the pipeline, and this author was unable to get it to work with the latest version of GATE (7.x). Therefore, a separate negation and possibility component was developed.

Rather than recreate NegEx and use its lists of specific expressions, we use lexical patterns to attempt to generalise the identification of negation and possibility. For example, a concept or finding preceded by a negating verb group (e.g. '*was not found*') or word (e.g. *no*, *absence* etc) within a certain window (e.g. between 0 and 3 intervening words) suggests that that concept is negated.

```
(
    (
        {VG.neg == "yes"}
        (
            TOKEN_WINDOW
            CONCEPT
        )[1, 5]
    ) |
    (
        CONCEPT
        {VG.neg == "yes"}
    )
):m
```

To ensure double negatives or negative possibility is not captured (e.g. '*does not exclude*'), negating phrases are only matched if they do not begin with 'not', as shown in the pattern below:

```
![“not”]
(
    (
        [“no|nor|any|deny|denie(s|d)|without|
        absen(t|ce)|exclude(d|s)|negative”]
    ) |
    (
        [“rule(s|d)?”]
        TOKEN_WINDOW
        [“out”]
    )
    (
        TOKEN_WINDOW
        CONCEPT
    )[1, 5]
)
```

where `TOKEN_WINDOW` is a flexible window of intervening words, and `CONCEPT` is a clinical term or event identified by a previous pipeline step (e.g. from MetaMap or another process; see Chapter 6). A similar pattern can be expressed to represent a negating expression following the concept. We use a similar approach to identify possibility via the presence of ‘hedge cues’[33] – words that indicate uncertainty or speculation:

```
(
    "possib(le|ility)|potential(ly)?|presum(e|ed|able|ably)|
    question(ed|able|ably)?|consistent|indicate(s|d)?|
    suggest(s|ed|ive)?|risk(s|ed)?"]
    (
        TOKEN_WINDOW
        CONCEPT
    )[1, 5]
)
```

5.2.6. Identifying quantitative and temporal concepts

The GATE `Tagger_Numbers`[9] component was used to identify and normalise spelt-out numbers and roman numerals to their arabic equivalents. This component does not handle ordinal numbers (21st, fourth etc), so a separate Gazetteer of ordinals from 1-31 (for day of the month identification) was created, e.g.

```
...
24th;val=24
twenty-fourth;val=24
twenty fourth;val=24
...
```

Gazetteer lists of units of measurement, their abbreviations and modifiers (in symbolic and text form e.g. '*less than*', '*at least*', <, >=) were created, and the output of these lookups were combined with JAPE patterns (e.g. Figure 5.4) to identify clinical relevant measurement concepts such as values and ranges of weight, volume (e.g. for drug dosages), length (e.g. for tumour sizes) and pressure.

Similarly, concepts of age, duration, frequency, and date/time were identified with JAPE string patterns combined with Gazetteers of month and day names, and relevant temporal units and their abbreviations. For example:

[9] http://gate.ac.uk/sale/tao/splitch21.html#sec:misc-creole:numbers

```
(
  {Number}
  {Lookup.majorType == time, Lookup.minorType == duration}
):expr
-->
  :expr.Duration = {value=:expr.Number.value, unit=:expr.Lookup.unit,
  period=:expr.Lookup.period, prefix=:expr.Lookup.prefix}

(
  {Duration}
  ("of")
  ["age"]
):expr
-->
  :expr.Age = {}
```

where Lookup entries for units of duration and their features are identified from the Gazetteer:

```
...
day;unit=H;period=24;prefix=P
days;unit=D;period=1;prefix=P
wk;unit=D;period=7;prefix=P
wks;unit=D;period=7;prefix=P
week;unit=D;period=7;prefix=P
weeks;unit=D;period=7;prefix=P
fortnight;unit=D;period=14;prefix=P
fortnights;unit=D;period=14;prefix=P
...
```

This allows duration values to be formalised according to the TimeML standard[32] (see Chapter 2). For example, '*for three weeks*' has a TimeML value of `P21D`, generated by mul-

tiplying the `value` and `period` features extracted by the above patterns and prepending and appending the `unit` and `prefix` features.

Frequency concepts are identified by similar patterns and gazetteers (e.g. *daily*, *weekly*, *once*, *twice*; for the full complement of patterns used in these modules, see the CD that accompanies this thesis), where the TimeML value is calculated by dividing the the `period` by the `value` features. Expressing singular 'day' concepts in hours allows frequency values to be calculated more accurately, e.g. '`three times a day`' → value=3, period=24, frequency value=24/3=8, TimeML value=`RP8H`; and '`twice daily`' → value=2, period=24, frequency value=24/2=12, TimeML value=`RP12H`.

TimeML defines a generic `TIMEX3` tag for duration, date, time and frequency concepts where each is distinguished by a '`type`' feature. The distinct annotations created for each in the first pass through the document are converted to `TIMEX3` annotations in a second pass (e.g. `Duration` → `TIMEX3.type=Duration`). In this second pass, number ranges or numbers preceded by a modifier in temporal expressions were given a '`mod`' feature as per the TimeML standard (e.g. '*no more than 3 days*' → `Duration.value="3", unit="D", mod="EQUAL_OR_LESS"`). These mappings were set up as follows: each Gazetteer entry has three features: positive context, one for negation, and pre-modifier 'or', for example:

```
more;pos=MORE_THAN;neg=EQUAL_OR_LESS;or=EQUAL_OR_MORE
earlier;pos=LESS_THAN;neg=EQUAL_OR_MORE;or=EQUAL_OR_LESS
```

These are matched by the corresponding patterns:

```
(
  {Lookup.majorType == value_modifier}
  {Number}
):mod
-->
  :mod.NumberModifier = {mod=:mod.Lookup.pos}
```

```
(
  ("no|not|never")
  {Lookup.majorType == value_modifier}
  ("than")?
  {Number}
):mod
-->
  :mod.NumberModifier = {mod=:mod.Lookup.neg}

(
  {Number}
  ("or")
  {Lookup.majorType == value_modifier}
):mod
-->
  :mod.NumberModifier = {mod=:mod.Lookup.or}
```

Identification and normalisation of date expressions

Although GATE includes a component for identifying and normalising date values, it does not identify abbreviated dates as typically occur in clinical notes such as 'on 8/26' (i.e. 26 August) or handled relative dates such as '*on the third post-operative day*' or '*on the day before discharge*', a separate component was developed for this purpose, again using JAPE expressions. For example, for UK dates:

```
(
  (DAYOFMONTH):day
  (DATESEP)
  (MONTH):month
  (DATESEP)
  (YEAR):year
):dt
-->
  :dt.Date = {day=:day.Token.string,
  month=:month.Token.string, year=:year.Token.string}
```

For US dates:

```
(
  (MONTH):month
  (DATESEP)
  (DAYOFMONTH):day
  (DATESEP)
  (YEAR):year
):dt
```

where DATESEP = "/" or "-" and the following regular expressions identify month, day and year expressions:

```
MONTH = (0?[1-9]) | (1[0-2])
DAYOFMONTH = (0?[1-9]) | (1[0-9]) | (2[0-9]) | (3[0-1])
YEAR = ([1-2][0-9]{3}) | ([0-9]{2})
```

Shorter date expressions, on their own, are ambiguous (11-12 could represent a value range, 11 December or 12 November, depending on locale), so patterns for matching these require a preceding prepositions and fixed locale. For example, to match US mm/yy (e.g. 8/92) or US mm/dd or mm-dd (e.g 09-26):

```
("on|in|from|before|after|during")
(
  (MONTH):month
  (DATESEP)
  (YEAR):year
):dt

("on|in|from|before|after|during")
(
  (MONTH):month
  (DATESEP)
  (DAYOFMONTH):day
):dt
```

For relative dates, such as '*2 days before admission*':

```
(
    ({Duration}):dur
    ({TemporalRelation})?:rel
    ({Event}):evt
):dt
-->
    :dt.Date-Rel = {rel=:rel.TemporalRelation.type, mod=:dur.Duration.mod,
    period=:dur.Duration.period, value=:dur.Duration.value,
    interval=:dur.Duration.unit, event=:evt.Event@string}
```

and for dates relative to a nonspecific event (e.g. '*three days ago*'):

```
(
    ({Duration}):dur
    ({TemporalRelation}):rel
):dt
-->
    :dt.Date-Rel = {rel=:rel.TemporalRelation.type, mod=:dur.Duration.mod,
    period=:dur.Duration.period, value=:dur.Duration.value,
    interval=:dur.Duration.unit}
```

In clinical discharge summaries, admission and discharge dates should be explicitly identified, either in a separate field in the EHR or in a clearly identifiable heading in the text. The framework's date handling component stores these dates as document-level features, to allow normalisation of relative and abbreviated dates. For example, if we know that the admission date was 2011-12-16 and the discharge date was 2012-01-18, then short dates such as 12/26 can be normalised to 2011-12-26.

Similarly, given an expression such as '*three weeks after discharge*', which generates a `Duration` concept with value `P21D` (see above), methods from the Java `Calendar` class can be used to generate a date 21 days after 2012-01-18, i.e. 2012-02-08.

Anaphoric date and duration expressions, such as '*on that date*' and '*during that time*' are linked back to the most recent, fully specified date or duration earlier in the document.

5.3. Evaluation

5.3.1. Text segmentation and MetaMap integration

The MetaMap integration component and the effect of its various configurations on performance were evaluated against two datasets: the 106kB, 11,000 word recommendations section of the *ESC European Guidelines on Cardiovascular Disease Prevention in Clinical Practice* in UTF-8 XML format, and 890 discharge summary documents in plain ASCII text format from the 2007 i2b2 corpus[34]. For each dataset, MetaMap was configured to return only SNOMED CT mappings, and the output of MetaMap's default text segmen-

tation – i.e. blank line delimited chunks – was used as the baseline, reference standard and this was compared against alternative text segmentation approaches: term phrase chunking (noun-, prepositional-, and adjectival-phrase chunks as described in Section 5.2.2), sentence chunking, and, for the clinical guidelines, XML element chunking (taking the contents of the source data's paragraph, list item and heading elements). For each, precision, recall and F_1-measure were calculated against the reference standard (see Chapter 4 for details of these evaluation metrics). The output of the verb group predicate patterns (see Section 5.2.4) were not used as input.

For the clinical guidelines data, tests were run both without and with MetaMap's `-term_processing option`, which treats the input as a single phrase for direct lookup into the UMLS Metathesaurus, and provides a mechanism for mapping composite phrases to a single identifier in UMLS.

Discharge summary data was provided in UTF-8 XML format. As the data contained only a single XML element for the text field, XML element chunking was not used. Documents varied in size from 30 to over 2000 words.

5.3.2. Events, negation and possibility in clinical notes

The performance of event boundary detection, type (e.g. `Treatment`, `Problem`, `Occurrence`), polarity (negation: either `POS` for positive or `NEG` for negated events), modality (possibility: either `FACTUAL` for events determined be true, `POSSIBLE` for events that may or may not be true, `PROPOSED` for planned events), was evaluated against a manually annotated corpus of 120 discharge summaries provided by i2b2 for their 2012 Natural Language Processing Challenge on temporal relations[10]. However, the main goal was evaluation of the negation module's performance in terms of polarity and modality assignment. Unannotated test data was provided by i2b2 in UTF-8 XML format and a Python script provided by the challenge organisers to evaluate the system output against the manually annotated gold standard.

[10] https://www.i2b2.org/NLP/TemporalRelations/Main.php

5.3.3. Temporal, quantitative and process concepts in clinical notes and guideline documents

The accuracy of temporal concept boundary detection, type, TimeML value and modifier was also evaluated against the manually annotated corpus of 120 discharge summaries provided by i2b2 for their 2012 Natural Language Processing Challenge on temporal relations. The manual annotations included both fully specified and relative date, durations and frequency concepts with the normalised TimeML value and modifier stored as features on each. As the data originate from US healthcare providers, dates were in US format, so the pattern for identifying UK dates (dd-mm-yyyy) was disabled.

Formative evaluation of the quantitative and process components in the framework was carried out by visual inspection of the output on a number of clinical guideline documents and anonymised discharge summaries to confirm that these contextual concepts were being correctly identified (see Figures 5.6–5.8 in Results).

5.4. Results

5.4.1. Text segmentation and MetaMap integration

Table 5.1 shows the time taken to annotate the guideline following default chunking compared with annotation of individual XML elements, sentences and phrases. Table 5.2 shows recall, precision and F-measure scores for MetaMap annotations produced from the different lexical units both with and without term processing. The output of the default-chunked input was used as the reference standard.

Table 5.1.: Clinical guideline processing times for different chunking approaches

	Default	Element	Sentence	Phrase
Time (s) w/o term processing	745	208	213	268
Time (s) w/term processing	$> 10^{5\dagger}$	$> 10^{5\dagger}$	$> 10^{5\dagger}$	325

† Process aborted after 3 hours with no output

As shown in Table 5.1, only phrase chunking allowed MetaMap's term processing to be used without excessive processing time (for the other chunking methods, processing

Table 5.2.: Clinical guideline recall/precision for Element, Sentence, Phrase (B) vs default chunking (A)

Input chunk	Match	Only A	Only B	Overlap	Rec.	Prec.	F_1
Element	4122	349	154	9	0.92	0.96	0.94
Sentence	4393	27	13	9	0.99	1.00	0.99
Phrase	4168	128	53	184	0.93	0.95	0.94
Phrase†	3889	224	118	367	0.87	0.89	0.88

† System run with MetaMap `-term_processing` enabled

was aborted after 3 hours with no output). Term processing did not make a substantial difference in processing time for phrase chunking, but caused the processing time for the other chunking approaches to increase dramatically. As show in Table 5.2, sentence chunking provides the most accurate output ($F_1 = 0.99$) relative to default chunking, however term processing is not possible with this method. Phrase chunking with term processing causes a significant drop in accuracy as quite different mappings are created than when default chunking is used without this option enabled; the reasons for this is discussed in Section 5.5.

As shown in Table 5.3, an attempt to process the entire discharge summary corpus using default and sentence chunking did not complete in a realistic time, and so processing was aborted after 74 documents had been processed in 6 hours (the reason for this failure is discussed in Section 5.5). Phrase chunking was also slow, but did complete in 17000 seconds (around 4 1/2 hours) for the 890 documents, generating 173,000 annotations, which gives an annotation rate of 10 per second.

Table 5.4 shows recall, precision, and F-measure scores for the MetaMap annotations produced from sentence and phrase chunks for the 74 documents for which processing completed from the output of all three chunking methods, taking the MetaMap output of default chunking reference standard as before. As with the clinical guidelines data, sentence chunking provides the most accurate output relative to default chunking, and both sentence and phrase chunking perform with similar accuracy on clinical discharge summaries as with clinical guidelines data ($F_1 = 0.99$ and 0.94 respectively). For the 74 documents, phrase-chunk processing generated 10,100 annotations in 840 seconds, giving

an annotation rate of 12 per second.

Table 5.3.: Discharge summary corpus processing times for different chunking approaches

	Default	Sentence	Phrase
Time (s)	$> 2 \times 10^{5\dagger}$	$> 2 \times 10^{5\dagger}$	17,100

† Aborted after 6 hours taken to process < 1/10 of the corpus

Table 5.4.: Discharge summary recall/precision for Sentence and Phrase (B) vs default chunking (A)

Input chunk	Match	Only A	Only B	Overlap	Rec.	Prec.	F_1
Sentence	2691	1	5	17	0.99	0.99	0.99
Phrase	2478	155	34	76	0.91	0.96	0.94

5.4.2. Events, negation and possibility in clinical notes

Table 5.5 shows the document macro-averaged and corpus micro-averaged precision, recall and F_1-measure scores for system-generated `EVENT` extents (boundary detection), and `Type` (`Problem`, `Test`, `Treatment`, or `Occurrence`[11]), `Polarity` (negative or positive) and `Modality` (possibility) F_1-measure scores for feature assignment, for the 120 discharge documents from the 2012 i2b2 corpus.

As shown in the table, corpus micro-averaged F_1-measure scores for `Type`, `Polarity` and `Modality` are significantly lower than the document macro-averaged scores as a result of the different way each is calculated (see Chapter 4). The macro scores show the average score per document for these features, given system events whose extents match the gold standard, whereas the micro scores show the average score over the whole corpus, taking into account false positives and false negatives. In other words, for a given system-generated `EVENT` annotation that matches or overlaps a gold-standard `EVENT` annotation, the accuracy (as measured by F_1) of negation and possibility assignment is 93% and 94% respectively, whereas these are reduced to 57% and 58% as a result of the number of false negatives (recall was only 62%, so 38% of all events were missed) and false positives

[11] a verb group

(precision was 82%, so 18% of text strings classified as events were falsely identified as such).

Table 5.5.: Identification of events, negation and possibility: macro- and micro-averaged metrics over 120 discharge summaries

Method	Precision	Recall	F_1	Type	Polarity	Modality
Macro	0.82	0.63	0.71	0.84	0.93	0.94
Micro	0.82	0.62	0.70	0.51	0.57	0.58

5.4.3. Temporal, quantitative and process concepts in clinical notes and guideline documents

Figure 5.5 shows example output generated by the temporal expression identification component on an anonymised discharge summary from the i2b2 2012 corpus. The pop-up box (*right of the figure*) demonstrates how the expression '*six years ago*' has been identified by the `MatchRelativeDatePost` rule (described at the end of Section 5.2.6) and normalised to 2005-02-08 by subtraction of 6 years from the admission date 2011-02-08 (*left of the figure*; NB admission date has been randomised).

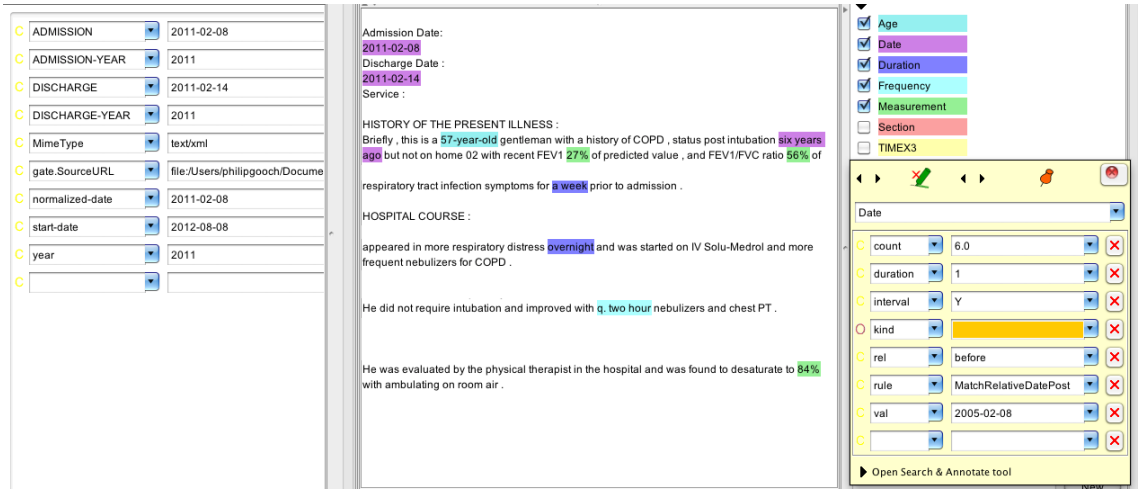


Figure 5.5.: Pattern-based identification of temporal expressions

Table 5.6 shows the document macro-averaged and corpus micro-averaged precision, recall and F_1-measure scores for system-generated TIMEX3 extents (boundary detection), and `Type`, `Val` and `Mod` F_1-measure scores for feature assignment, for the 120 discharge

documents from the 2012 i2b2 corpus. The `Type` represents TIMEX3 type assignment accuracy (i.e. `Duration`, `Date`, `Time`, or `Frequency`) . The `Val` score represents accuracy of TimeML value calculation. This takes into account unit conversion, e.g. a value of `P36H` in the system output will score as a match against a value of `P1.5D` in the gold-standard output (provided the concept extents and `Type` also match). The score for the `Mod` feature represents accuracy of the TimeML modifier (`NA` for specific values, `LESS`, `MORE`, `APPROX`, `START`, `END` and `MIDDLE` for concepts preceded by an appropriate modifier).

As shown in the table, corpus micro-averaged F_1-measure scores for `Type`, `Val` and `Mod` are significantly lower than the document macro-averaged scores as discussed above. These differences provide that, for a given document, the accuracy of type and normalised value will be 90% and 78%, respectively, for a given temporal annotation identified by the system that matches the same annotation in the gold standard, whereas over the corpus as a whole, these accuracies are reduced to 68% and 59%, due to the effect of false negative and false positive system TIMEX3 extents.

Table 5.6.: Temporal concept identification: macro- and micro-averaged metrics over 120 discharge summaries

Method	Precision	Recall	F_1	Type	Val	Mod
Macro	0.85	0.80	0.81	0.90	0.78	0.88
Micro	0.83	0.77	0.80	0.68	0.59	0.68

Formative evaluation results

Figure 5.6 shows an extract from a discharge summary with temporal, quantitative and process information that has been identified by the patterns described in Sections 5.2.6 and 5.2.4 highlighted.

Figure 5.7 shows a guideline recommendation identified for the ‘patients with diabetes’ population, with quantitative concepts identified by the patterns described in Section 5.2.6 highlighted.

Figure 5.8 shows a clinical action identified for the ‘patients with established cardiovascular disease’ population, if aspirin is contraindicated.

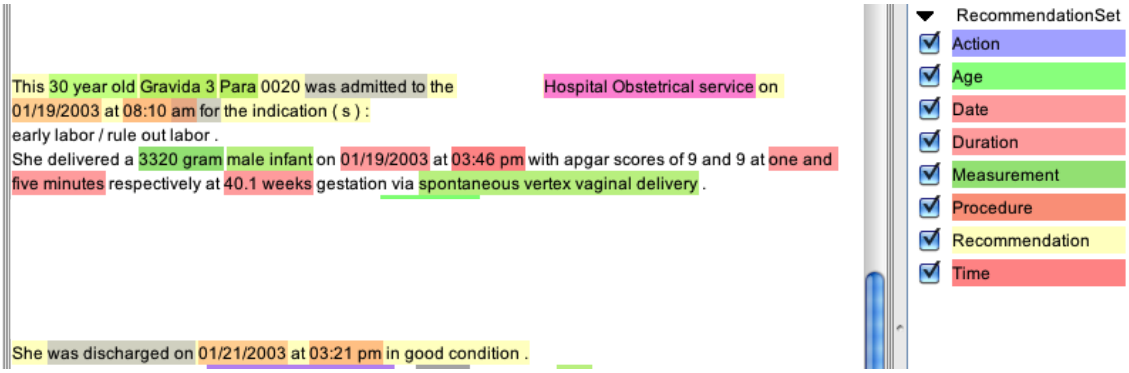


Figure 5.6.: Temporal and quantitative concepts identified in an anonymised clinical discharge summary
Date and Duration concepts shown in orange-red. Clinical process information 'was admitted to' and 'was discharged on' shown in grey.

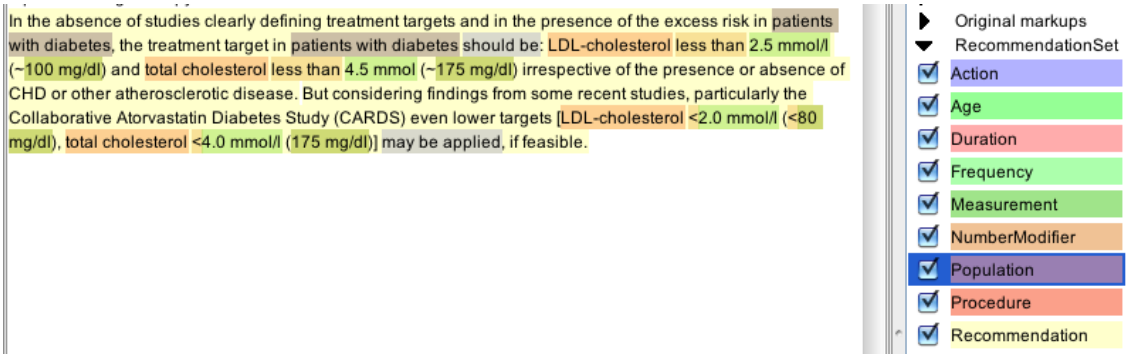


Figure 5.7.: Quantitative concepts identified in a clinical guideline
Quantitative concepts shown in green; modifiers shown in light orange. Population groups shown in light brown.

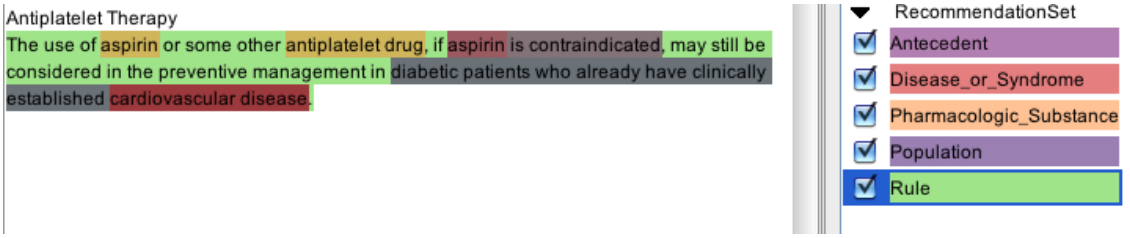


Figure 5.8.: IF...THEN rule identification in a clinical guideline
Antecedent shown in mauve; disease concepts shown in red; treatment shown in orange; population group shown in grey.

5.5. Error analysis and discussion

5.5.1. Text segmentation and MetaMap integration

The clinical discharge summaries contained physical examination and test result information in list format without punctuation; the inability of MetaMap to adequately process such text without syntactic structure is a known problem[10]. However, with phrase chunking, processing rate decreased only slightly (from 12 per second to 10 per second) as the number of documents increased fro 74 to 890. This suggests that when processing by phrase, the time taken will increase roughly linearly with the size of the document, but seems to be independent of the complexity of the underlying text. For default or by-sentence processing, however, time taken appears to be dependent on both document size and complexity.

The reduction in recall and precision (of between 4-9%) of phrase-by-phrase processing on both the clinical guideline and the discharge summaries was largely caused by verb group predicates that are identified by the VG chunker (see Section 5.2.4) but were not submitted to MetaMap (only the output of the candidate term phrase chunker was used for concept mapping, so unattached predicates were discarded). However, these verb groups contained SNOMED CT mappings, for example, '*transferred*', '*evaluated*', and '*discharge*', which were picked up from default and sentence chunking. Although we are interested in these predicates for identifying events and process information, arguably it is not necessary to map these individual process phrases to terminology concepts, so this reduction in accuracy may be acceptable. Alternatively, of course, these predicates can be added to the set of input annotations to the MetaMap pipeline component.

Overall, processing by sentence provided the best trade-off between speed and accuracy. However, phrase chunking allows rapid term extraction and mapping from input without syntactical structure, such as bulleted lists, which can require many hours of computation if processed directly by MetaMap[10]. Also, as shown in Table 5.1, processing by phrase processing allows the `-term_processing` option to be used. This is useful as it allows composite phrases in the guideline to be mapped to their pre-coordinated terms in UMLS,

rather than to multiple terms requiring post-coordination. Some examples are shown in Table 5.7.

Table 5.7.: Clinical guideline phrase chunking mappings both without and with (bold) term processing

Phrase	Mappings
self-monitoring of blood glucose	Self-monitoring(C0588436); Glucose measurement, blood(C0392201)
	Self-monitoring of blood glucose(C0005803)
quality of life	Quality(C0332306); Household composition(C0595998)
	Quality of life(C0518214)
management of risk factors	Administration(C0001554); History of - risk factor(C0455624)
	Risk management(C0035649)
foot pulses	Entire foot(C1281587); Pulse(C0391850)
	Pedal pulse(C0232157)
increases in blood pressure	Increase(C0442805); Blood pressure finding(C1271104)
	Elevated blood pressure(C0497247)

As shown in Table 5.7, term processing produces quite different mappings than when this option is not used. These differences explain the lower recall and precision when this option was used (0.87 and 0.89 with term processing vs 0.93 and 0.95 without). The mappings with term processing seem to provide a single, more useful, mapping for each composite phrase. '*Quality of life*' is more usefully mapped to the single, eponymous `Finding` concept, rather than to a '*Quality*' `QualitativeConcept` and a '*Household composition*' `Finding`. Similarly, '*self-monitoring of blood glucose*' is possibly better represented by the eponymous `DiagnosticProcedure` than by post-coordinating a generic `TherapeuticOrPreventativeProcedure` '*self-monitoring*' concept and a `LaboratoryProcedure` '*Glucose measurement, blood*' concept. So, although accuracy against the nominal reference standard output is reduced, in practice, the annotations may be more clinically useful.

5.5.2. Event detection, negation and possibility

The low recall in event detection was largely a result of abbreviations in the discharge summaries (e.g. '*benzos*', '*the Rita*'), trade names for drugs ('*Klonopin*') not picked up by the concept recogniser. Reduced precision resulted from identifying generic verb groups as event occurrences. However, as noted in the Section 5.3, the main goal of this part of the evaluation was performance of negation and possibility assignment. If we take the macro scores in order to consider the performance of this component on its own, then the simple lexical patterns described in Section 5.2.5 appear to have performed well, giving F_1 measures of 0.93 and 0.94 for negation and possibility assignment respectively. This suggests that, on the evaluation corpus at least, simple patterns, rather than specific, hard-coded expressions as used by NegEx, perform well – Chapman et al[22] cited precision and recall of 84.5% and 77.8% for negation detection, giving an F_1 measure of 0.81, although they used a different corpus of discharge summaries than the one used here for evaluation. A simplification of NegEx based on the presence of negating words within a flexible window of the target term was also proposed by Koeling et al.[35], although only negating words preceding the term were considered, whereas here we have used negating expressions both preceding and following the target term.

Instances where the simple negation detection patterns fail, however, include expressions such as '*gram negative bacteria*, and '*she didn't know why she is HIV positive*'. Future work might look at making the simple patterns a little more sophisticated without having to create a fully specified list of complete expressions.

5.5.3. Temporal and process concepts

In terms of automatically identifying temporal concepts and formalising them into TimeML expressions, there has been little previously reported research on evaluation of methods to achieve this, particularly in the clinical domain. Chang and Manning[36] have recently reported on SUTime, a rule-based system that is part of the Stanford CoreNLP framework[12] that uses similar temporal morphemes and composition rules as those presented here as

[12] http://nlp.stanford.edu/software/corenlp.shtml

JAPE expressions. They reported precision, recall and F-measure scores of 0.88, 0.96 and 0.92 for extents, and TimeML `Type` and `Value` F-measures of 0.96 and 0.82, against a corpus of general newswire texts derived from the TimeBank corpus[37]. Although nominally superior to the results presented here for the clinical discharge summaries, their evaluation corpus is not comparable to the corpus of discharge summaries used in this work. Also, it is not clear whether the authors report micro- or macro-averaged figures.

Also, as previously noted, working with patient notes presents particular difficulties for NLP applications, such as use of non-standard acronyms and abbreviated expressions. Although the performance of SUTime has not been formally evaluated against the i2b2 corpus, running the online demo[13] of SUTime against a small selection of discharge summaries suggests that it does not recognise abbreviated date expressions such as '*on 10/19*', abbreviated durations such as '*15 min*' nor dosage frequency abbreviations such as '*t.i.d*', and also annotates `Age` concepts (e.g. '*a 48 year old man*') as `Duration`.

Strötgen and Gertz[38] recently reported on HeidelTime, another rule-based temporal tagger developed to identify temporal expressions in four different domains: narrative, colloquial, news, and biomedical. In the biomedical corpus that they created for evaluation, they reported precision, recall and F-measure scores of 0.95, 0.66 and 0.78 for extents and TimeML `Value` F-measure of 0.70. Again, their evaluation corpus is not comparable to the one used in the present work, although it is at least in the same domain. Unlike SUTime, HeidelTime is available as a standalone Java component, which should allow it to be integrated into the framework via the GATE API (in a similar way that the MetaMap server was integrated, as described in Section 5.2.3). Future work could compare the performance of HeidelTime against that of the current component on the i2b2 corpus.

To identify the causes of the errors made by the GATE temporal expression component developed here, 20 documents with the lowest scores in precision, recall or `Value` F-measure were selected from the corpus and the discrepancies between the system output and gold standard analysed. Errors are summarised in Table 5.8, in the examples given, the left-hand-side expressions are from the gold standard.

[13] http://nlp.stanford.edu:8080/sutime/process

Table 5.8.: Analysis of errors in temporal expression identification and formalisation

Error type	Examples	n
Incorrect relative date `Value` calculation	postoperative day number two: 2009-08-26 *vs* 2009-08-19; hospital day # 1: 1992-09-21 vs 1992-09-22; the Sunday prior to admission: 2016-03-13 *vs* <null>	50
Missing abbreviated event date/duration or other temporal abbreviation	POD#6; last couple of days; on the 16th; stent [05-26]$_{\text{Date}}$; day of life #1; through [12-21]$_{\text{Date}}$	38
Missing 'orphaned' frequency, duration or time expressions	[five]$_{\text{Frequency}}$ grafts; days #5 and [6]$_{\text{Date}}$; [1]$_{\text{Time}}$ and 5 minutes; [four]$_{\text{Frequency}}$ past hospitalizations;	5
Incorrect type	for [ten days]$_{\text{Duration}}$ after discharge *vs* [ten days after discharge]$_{\text{Date}}$; for [4 days]$_{\text{Duration}}$ prior to discharge *vs* [4 days prior to discharge]$_{\text{Date}}$; [5 hours of life]$_{\text{Time}}$ *vs* [5 hours]$_{\text{Duration}}$ of life; [three days ago]$_{\text{Date}}$ *vs* [three days]$_{\text{Duration}}$ ago; [the three days]$_{\text{Duration}}$ prior to admission *vs* the [three days prior to admission]$_{\text{Date}}$	29

As shown in Table 5.8, incorrect `Value` calculation of relative dates ($n = 50$) and abbreviated event dates and durations ($n = 38$) form the bulk of the errors identified in the 20 documents sampled. Problems with relative date value calculation stem from incorrect identification of the antecedent source date. Such dates are often not explicit in the document. For example, calculation of the correct value for '*postoperative day number two*' requires correct identification both of the surgical event (which may be expressed in many ways, such as '*went for surgery*', '*was transferred to the operating room*' etc), the date that this event occurred, and then recognition that '*postoperative*' refers to a time after this date.

In the present component, calculation of relative date values is limited to dates relative to the date of admission or discharge in the current component. For example, '*the next day*' in the absence of a prepositional attachment to an admission or discharge event, will, by

default, be calculated as the day following admission. However, if the text has '*pt was sent to the ICU 20/12/2002. He received a dose of furosemide and was transferred to the ward the next day.*' then the relative date calculated will be incorrect. One potential solution would be to simply link relative dates back to the most recently mentioned date. However, if the most recently mentioned date is a historical episode, such as '*pt was diagnosed with CHF in 8/04*' then a later mention of '*the next day*' is more likely to be relative to some other date or period in the current episode than the immediately preceding, historical one. Clearly, identification and calculation of relative date values in clinical notes required more sophisticated handling than linking them to either the fixed dates (admission and discharge) or the most recently mentioned date – although this is also a weakness shared by other, general temporal expression parsers such as SUTime[36].

Difficulty distinguishing relative dates from durations was also a common problem ($n = 29$ in the 20 documents sampled). This was partly a result of inconsistent annotation in the gold standard: for example, instances of '*several months ago*' being annotated as an approximate `Date` (with features val="2004-06" mod="APPROX") vs '*three days ago*' being annotated as `Duration`. There may be some mileage making use of the preceding preposition to disambiguate the two, e.g. '*over the three days prior to admission*' would be a `Duration` but the same expression without the preceding preposition, '*three days prior to admission*', would be a `Date`.

A less frequently encountered error ($n = 5$) was the failure to pick up 'orphaned' temporal expressions, i.e. those linked by a conjunction to an earlier or later, more fully specified expression – for example '5' in '*day 4 and 5*' or a quantification of an event functioning as the frequency that that event occurred (e.g. 'five' in '*five previous operations*').

In more formally written clinical texts, such as guidelines, the temporal concept identification component may be more useful than its current performance on discharge summaries suggests, although it has not yet been evaluated against guideline texts. However, figures 5.6 to 5.8 show how the framework's temporal, quantitative and process concept identification components can highlight information that may be useful for identifying IF...THEN statements[39], clinical goal phrase patterns (see Chapter 2, Section 2.3.4), or

individual data items required for guideline formalisation (Y. Shahar, personal communication, 6 July 2011). This may assist in the transformation of guideline statements manually extracted into semi-formal representation into a fully structured, formal representation[40] . For example, given the guideline statement:

Doxycycline: 100 mg orally twice a day for 7-9 days.

the pipeline produces the following output in XML format:

```
<Paragraph>
  <Drug>Doxycycline</Drug>:
  <Measurement unit="mg" value="100.0" type="weight">100 mg</Measurement>
  <Procedure>orally</Procedure>
  <Frequency value="2" period="24" unit="H">twice a day</Frequency> for
  <Duration mod="APPROX" unit="D"
    value-high="9.0" value-low="7.0">7-9 days</Duration>.
</Paragraph>
```

and this can be transformed using an XSLT transformation to an equivalent statement in the Asbru formalism:

```
<plan name="Doxycycline: 100 mg orally twice a day for 7-9 days">
  <cyclical_plan>
    <frequency value="12" unit="H"/>
  </cyclical_plan>
  <duration>
    <min value="7" unit="D"/>
    <max value="9" unit="D"/>
  </duration>
</plan>
```

Similarly, for the guideline statements:

> For patients with proteinuria with at most 1 gram per 24 hours, the blood pressure should be controlled to below 130/85 mmHg.
>
> For patients with proteinuria in excess of 1 gram per 24 hours, blood pressure should be controlled to 125/75 mmHg.

the pipeline produces the following XML output:

```
<Paragraph>
  For patients with
  <DiseaseOrSyndrome>proteinuria</DiseaseOrSyndrome>
  with at most
  <Measurement unit="g" value="1.0" type="weight"
    mod="EQUAL_OR_LESS">1 gram</Measurement>
  <Frequency period="24" value="1" unit="H">per
    <Duration period="1" value="24" unit="H">24 hours</Duration>
  </Frequency>,
  <Test>the blood pressure</Test>
  should be controlled to below
  <Measurement unit="mmHg" type="pressure"
    mod="LESS_THAN">130/85 mmHg</Measurement>.
</Paragraph>

<Paragraph>
  For patients with
  <DiseaseOrSyndrome>proteinuria</DiseaseOrSyndrome>
  in excess of
  <Measurement unit="g" value="1.0" type="weight"
    mod="MORE_THAN">1 gram</Measurement>
  <Frequency period="24" value="1" unit="H">per
    <Duration period="1" value="24" unit="H">24 hours</Duration>
  </Frequency>,
  <Test>blood pressure</Test>
  should be controlled to
  <Measurement unit="mmHg" type="pressure">125/75 mmHg</Measurement>.
</Paragraph>
```

which could similarly be transformed into statements and data item definitions in Asbru

or other formal guideline modelling formalisms, although mappings and transformation rules for doing so remains an area for future research (see Chapter 9).

5.6. Summary

In this chapter, a number of core components in the clinical information extraction framework have been developed and evaluated, focusing on the integration of MetaMap with GATE to identify and classify clinical concepts, and extraction of quantitative and temporal information. As the components purely utilise lexico-syntactic patterns, they do not require training on different data sets, and the patterns are easy to extend. Parsing, normalising and reasoning with temporal expressions in free text is a complex area of research in its own right: we have only touched on it briefly for the purposes of evaluating the component developed here. Although event, temporal concept and negation detection were the only framework components quantitatively evaluated against a gold standard corpus in this chapter, all these components are utilised as an ensemble in Chapter 8 and evaluated against other gold standard, curated corpora.

For documents containing properly structured text, such as guideline recommendations, easily parseable into phrases by MetaMap, submitting individual sentences provides the best tradeoff between speed and accuracy in comparison to attempting to process blank-line delimited segments (i.e. paragraphs) in one go. However, for large documents, or those that contain unstructured or complex phrases, pre-processing the document into candidate term phrases provides a useful method for dramatically reducing the time required by MetaMap to map the text to UMLS concepts, although annotation accuracy is slightly reduced.

One limitation of the text segmentation method presented here for identifying candidate phrases is that it does not provide a way of identifying the potential classification of the candidate term (step 2 of the Krauthammer & Nenadic 3-step process), leaving this up to MetaMap. In the following Chapter 6, we consider an alternative, lightweight approach to clinical concept recognition and classification that leads to significant performance improvements over MetaMap both in terms of speed *and* annotation accuracy.

6. Simplifying concept identification in clinical narratives: semantic decomposition of ontology resources for creating term recognisers

6.1. Introduction[1]

In Chapters 2 to 4 we saw how ontologies such as the Foundational Model of Anatomy (FMA)[2], terminologies such as SNOMED CT, and compendia such as the UMLS that aim to integrate these resources into a comprehensive vocabulary, have formed the core knowledge bases for mapping text strings to concepts in the clinical domain. In Chapter 5, we saw how MetaMap (and its now deprecated, standalone version, MMTx) has formed a key component of many systems for identifying UMLS concepts in unstructured text, and we evaluated some approaches to improving its performance (in terms of processing speed and more focused term mapping) when processing large documents or text lacking syntactic structure. These approaches relied on reducing the search space over which MetaMap needs to consider the possible mappings of individual words in the phrase, and the complete phrase, to terms in the UMLS. However, apart from some edge cases, the processing speed improvements were fairly modest (although the process always completed), as complete noun phrases, prepositional phrases and verb phrases still needed to be submitted for term identification, semantic type assignment and terminological map-

[1]Some of the results presented in this chapter have been published in 'Systematic identification and correction of spelling errors in the Foundational Model of Anatomy'[1].

ping. In this chapter, we describe and evaluate a method for creating efficient concept recognisers through morpheme-based decomposition of ontologies and a simple grammar for identifying potential terms.

MetaMap adds value in the form of its comprehensive coverage[3] and additional metadata that it adds to the concepts that it identifies, such as the Concept Unique Identifier (CUI) to enable interoperability, the UMLS semantic type, the canonical name (the UMLS Preferred Name, so the matched phrase may be a synonym of this), and the individual resources in which it has located the term. But MetaMap is a fairly heavyweight tool, requiring 10GB of disk space to install and 4GB RAM to run. If we had a rapid, accurate method for extracting a term and its semantic type directly from one or more ontologies, then it would be possible to obtain the CUI and other UMLS metadata directly from the UMLS Metathesaurus, or, if further mappings are required, from a more focused MetaMap search.

At the simplest level, this could be achieved by using the ontology as a large gazetteer – a lookup list. But this is inefficient for a number of reasons:

1. Ontologies can be very large in their native form, for the example, the FMA is over 200MB and comprises over 150,000 terms; SNOMED CT comprises over 315,000 terms.

2. Ontologies may not be complete: how does one identify terms that 'should' be in the ontology?

3. Ontologies may not contain all term synonyms or different lexical variants

A number of solutions have been developed to address these problems. Tools such as the National Library of Medicine's Lexical Variant Generation (LVG) tools[4] can be used to pre-process and normalise text prior to matching it against ontology terms. The innovative Textpresso ontology[5] addresses the lexical variant problem by including, for each ontology term, a regular expression that will help identify that term in free text, for example '[Ee]mbryos?' This approach works well for single word expressions but becomes

unwieldy for multi-word expressions as the number of regular expression combinations increases.

Another approach involves identifying domain-specific features that can be leveraged to identify potential words that are likely to form domain ontology terms. It has long been recognised that biomedical and clinical terms are highly compositional, being made up of well-defined linguistic fragments known as *morphemes*. A morpheme is the smallest linguistic unit that has semantic meaning, which may be a word (free morpheme) or a word fragment (bound morpheme) such as a prefix, root, or suffix. In particular, biomedical terms are frequently composed of Latin and Greek morphemes. Terms composed of Latin and Greek morphemes are known as neoclassical compounds; the rules for joining them are known as combining forms. Free neoclassical morphemes include *cephalon*, *metacarpus*; bound morphemes *cirrh-*, *derm-* (roots) and *-itis*, *-rrhea* (suffixes). Analysis of neoclassical compounds can help identify and classify unknown terms[6]; for example disease terms might be identified via suffixes *-itis*, *-osis*, or *-opathy*. The National Library of Medicine's SPECIALIST Lexicon[4] includes tools for identifying neoclassical compounds with a mapping to their English meaning and whether they function as root, prefix or suffix[2], although they are not classified according to semantic type (e.g. whether they describe a disease state, an anatomical structure, a qualitative concept, a chemical structure). However, neoclassical combining forms on their own are insufficient for term identification, providing high precision, but low recall[7].

Alternatively, ontology terms can be 'learnt' directly from text, via linguistic patterns, or graph-based machine-learning techniques[8][9]. Most well-known of the former approaches are the 'Hearst patterns'[10]: lexico-syntactic patterns that imply hypernym–hyponym (class membership or classification) relations between noun phrases. For example:

> Bruises, cuts, *and other* injuries ⇒ bruise `is_a` injury, cut `is_a` injury
>
> Diseases *such as* atherosclerosis ⇒ atherosclerosis `is_a` disease

Such patterns, as with neoclassical combining forms, identify potential terms with high precision, but with low recall[11]. The patterns can be augmented via bootstrapping

[2] `http://www.ncbi.nlm.nih.gov/books/NBK9680/`

(where unknown words can be classified by their relation to known terms in a phrase), for example:

scaphoid, lunate, triquetral and pisiform ...

– if we know that the *scaphoid* and *lunate* are bones of the wrist, we can infer that *triquetral* and *pisiform* are also. Also, ontology properties and relations can be used to infer the likely classification of a candidate word. For example, if an ontology contains concepts such as `Disease_or_Syndrome` and `Pharmacologic_Substance`, joined by the relation *treats* (and its inverse *treated_by*, i.e. `Disease_or_Syndrome` *treated_by* `Pharmacologic_Substance` and `Pharmacologic_Substance` *treats* `Disease_or_Syndrome`), then from the phrase

vancomycin treatment for MRSA failed

we can potentially infer that vancomycin `is_a Pharmacologic_Substance` and MRSA `is_a Disease_or_Syndrome`.

These approaches have been shown to improve recall, but reduce precision[11], and in any case, these approaches are more often used to augment existing ontologies or to build ontologies from scratch, rather than identify concepts in free text using an existing ontology.

Recently, a tool has been developed to make direct ontology lookup more efficient: mGrep[12]. mGrep compiles ontology or dictionary terms into a compact, radix trie structure (see Figure 6.1 for an example). A radix trie allow edges to be labelled with sequences of characters, rather than a single character as per a regular trie[13]. This is particularly useful for efficient storage and retrieval of terms that share common prefixes or roots. This feature makes them a good choice for ontologies rich in neoclassical roots, such as anatomical and clinical terms. mGrep has been shown to perform with higher precision but lower recall than MetaMap when identifying anatomical terms against the FMA, general concepts from SNOMED-CT, and disease concepts from UMLS[14], with the benefit of greatly reduced processing time – two orders of magnitude – in comparison to MetaMap. However, mGrep is noted to have two limitations: 1) it does not generate lexical variants,

which implies that it can only make exact matches, and 2) it requires both the dictionary *and* input text to be pre-processed into a three-column tab-delimited format[14], details of which do not appear to be documented. Furthermore, semantic type assignment requires additional pre-processing[15].

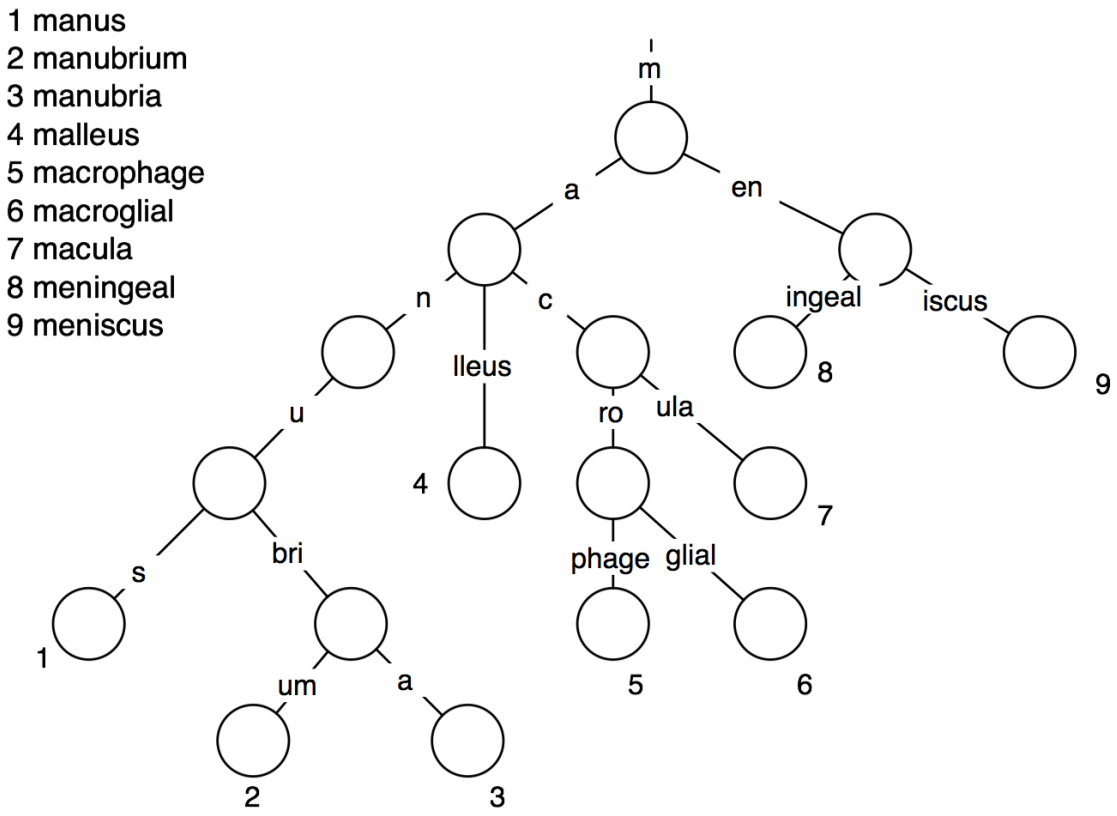


Figure 6.1.: Example radix tree representation of words from the Foundational Model of Anatomy

There is a need, therefore, for a systematic method for creating a concept recogniser from a given ontology without the overhead of looking up against the entire ontology, while allowing lexical variation in the input text without requiring it to be preprocessed. In this chapter, we introduce a method for semantic decomposition and recombination of ontology resources that addresses this problem. The idea of semantic decomposition of ontologies is not new, but previous approaches have considered the decomposition of the ontology logic to create reusable, logically independent modules[16]. Here, we are interested only in the ontology terms as dictionary entries and their corresponding semantic types, and decomposing multi-word expressions in the ontology into reusable morphemes that can be

recombined into candidate terms.

Tong et al.[17] decomposed the Gene Ontology into individual tokens and calculated the positional entropy of each token via the probability of token t appearing at position p in a given multi-word ontology term. However, this method was not applied to identifying potential ontology terms in free text. Still, the idea of morpheme-based term identification is not new either. As far back as 1994, Ananiadou[18] described a grammar of combination rules for neoclassical morphemes to recognise potential new terms. Neoclassical roots, prefixes and suffixes were extracted from immunological texts, and roots were classified as free (e.g. '*cyst*'), partially bound, either to a suffix or prefix (e.g. '*cyt-*' or '*-cyt*'), or requiring binding to a suffix (e.g. '*oo-*', '*or-*'). Similarly, morphemes were classified as occurring in both general and term usage (e.g. '*em-*' in 'emphasis' and 'embolism') or as indicative of a potential term only (e.g. '*leuk*'). However, no implementation or evaluation was provided at the time.

In this chapter, we describe in detail the semantic decomposition and rule-based morpheme recombination process given one or more ontology resources as a source dictionary, and apply it to identifying and classifying candidate terms in unstructured text. We demonstrate the approach using two ontologies: the Foundational Model of Anatomy (FMA)[2] and the Disease Ontology (DO)[19], and evaluate its performance on a small corpus of patient progress notes, surgery and radiology reports. Finally, the results are compared with the performance of both MetaMap and direct ontology lookup on the same corpus.

The FMA and DO were selected as Shah et al.[14] also considered recognition of anatomical and disease concepts when comparing mGrep with MetaMap. More importantly, identification of anatomical terms are central to the identification of the contexts and locations of other concepts, for example

- the location of disease, morbidity, or injury
- the location of symptoms, signs and findings
- the location of surgery, pathology and radiology procedures, and administration

routes of medication

As we shall see in Chapter 8, consideration of these contexts is important when identifying processes of care in the clinical narrative. Finally, applying the process to a second ontology – in this case the DO – was important in order to see if the semantic decomposition process can be generalised.

6.2. Method

Using ideas from Tong et al.[17] and Müller et al.[5], the method combines semantic decomposition of multi-word expressions with regular expressions to identify lexical variants. The method comprises four phases: 1) the token-centric decomposition phase; 2) the quality assurance phase; 3) the classification phrase; and 3) the recombination phase. This is described in the following section (see Figure 6.3 for an overview of the first three phases of the process as applied to the FMA).

6.2.1. Token-centric decomposition

1. Extract term names from the ontology. If the ontology is in the standard OWL format, this is done by taking the string value of each `rdfs:label` element, which can be obtained using regular expressions or via an XPath query. If the ontology is in a database, then terms can be extracted via the corresponding database query (SQL or SPARQL depending on the database type).

2. Tokenise each term using whitespace as the token delimiter to generate a line-break delimited list of tokens.

3. Deduplicate and sort the token list and remove stopwords (determiners, prepositions).

6.2.2. Quality assurance

This phase involves running a biomedical spell-checker on the deduplicated list of tokens extracted from the ontology terms. In this study, a GATE integration plugin was developed

in Java using the the National Library of Medicine's (NLM) SPECIALIST GSpell Spelling Suggestion Java API. GSpell uses a number of different algorithms to retrieve similar words to the input word or term from a user-defined dictionary (the NLM Specialist Lexicon is used by default) and returns the top N candidates based on edit distance from the input word. The algorithms include Metaphone (a phonetic-based measure that improves on Soundex), homophones, n-grams, bag-of-words, and a NLM lookup list of common misspellings.

The plugin was configured to ignore capitalised words, those beginning or ending with a digit, and those less than 2 characters long. Spelling suggestions outside an edit distance of 2 were ignored. For words identified as misspelt, the integration plugin stores spelling correction suggestions as an feature on each word (see Figure 6.2 for an example from FMA tokens). Each correction was manually reviewed by checking the suggested spelling against the online MedlinePlus medical dictionary and Google. Variations in US/UK spelling were ignored. Each misspelt word was then substituted for the corrected, consensus version of the word in the token list, and misspelt word in the original ontology is updated with the corrected version.

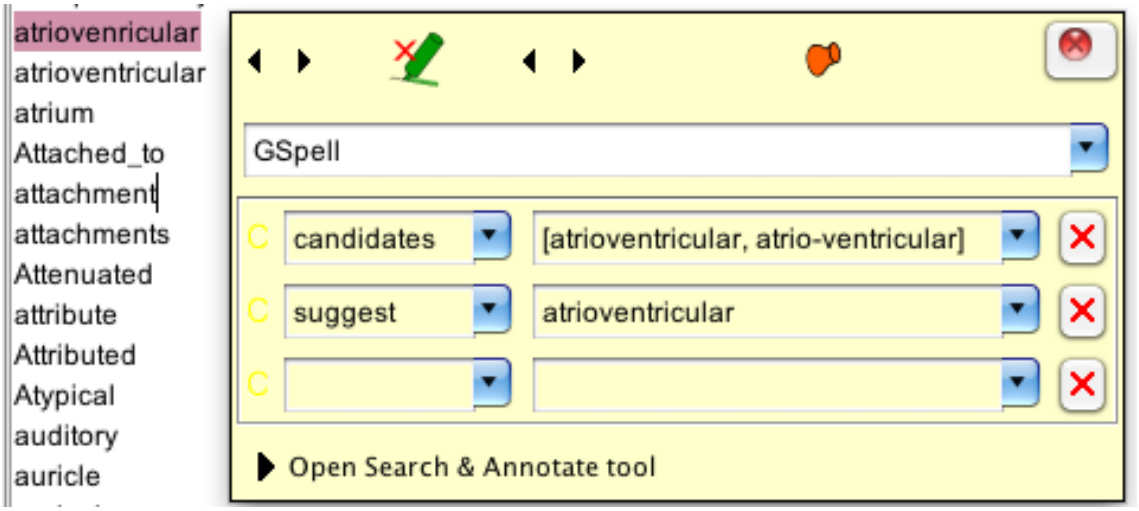


Figure 6.2.: Spelling error correction of words from the Foundational Model of Anatomy

6.2.3. Lexical and semantic classification

1. Classify each token according to its lexical type (noun, proper noun, adjective)

2. Lemmatise tokens into morphemes
3. If possible, classify each morpheme according to its semantic subtype from the ontology (e.g. part, space, substance)
4. Reduce each set of morphemes by identifying those sharing common roots and suffixes

6.2.4. Semantic recombination

Regular expressions are created over the union of entries (with morphological variants) in each set of classified morphemes. For example:

```
ont_nounPatt = ... macula | malleus | mandible | manubri(um|a) | manus ...
ont_adjPatt = ... hepatic | humeral | hyoid | ileal | iliac ...

neoclassicalSuffix = ... ineum | ionis | iores | ioris | iorium | iousus ...
neoclassicalPrefix = ... abdom | acanth |acetabul ...
```

For free morpheme patterns (whole words), we add word boundary constraints and allow for plurals:

```
ont_noun = \b( + ont_nounPatt + )?s\b
ont_adj = \b( + ont_adjPatt + )\b
```

For bound morpheme patterns, a starting boundary is specified for prefixes, with an open-ended closing boundary:

```
ont_prefix = \b( + neoclassicalPrefix + )
```

For suffixes, a configurable `minPrefixLength` parameter specifies how many characters must occur at the start of the word before the suffix:

```
ont_suffix = \b(\w{minPrefixLength,})( + neoclassicalSuffix + )\b
```

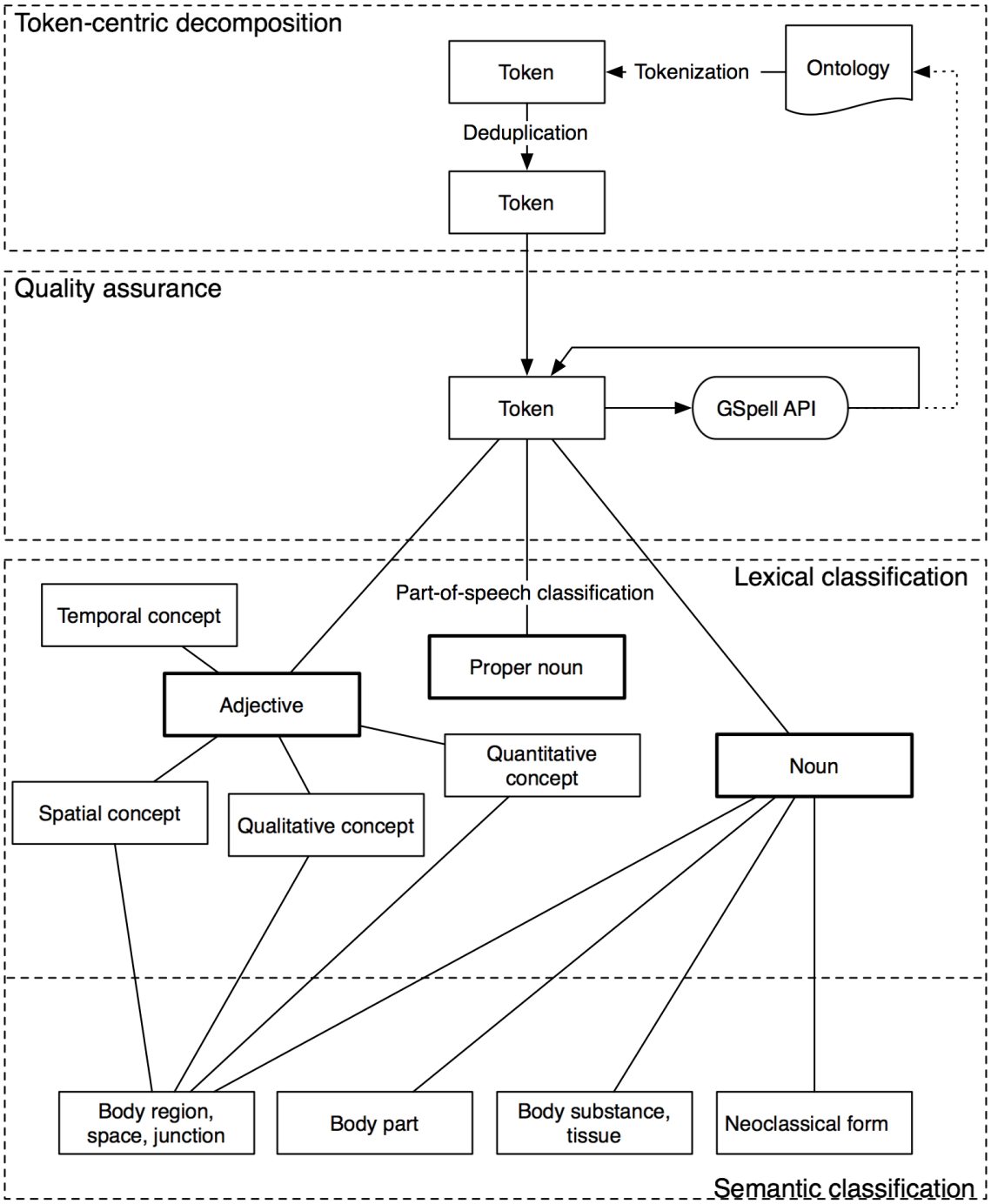


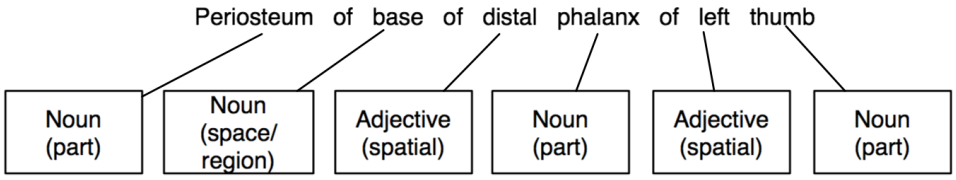
Figure 6.3.: Overview of semantic decomposition process as applied to the FMA Top: Tokenization phase; middle: QA phase; bottom: classification phase.

Recombination patterns are then applied over the regular expressions to identify candidate noun phrases and prepositional phrases that generalise the lexical and semantic structure of the original ontology terms. For example, noun phrases (NP) and prepo-

sitional phrases (PP) can be constructed in a similar way as described in Chapter 4, the difference here is that the constituent nouns and adjectives have been selected from ontology morphemes:

```
NP = ont_adj{0,5} (ont_suffix | ont_prefix | ont_noun | ont_properNoun){1,5}
PP = NP "of|on" NP
Term = NP | PP
```

These are generic patterns, independent of the underlying ontology classes. If morphemes have been classified according to semantic subtypes from the ontology – e.g. in the case of the FMA, body space/junction, body substance, body part/organ component – domain-aware patterns can be constructed in a similar way (see Figure 6.4)



Generic:

```
NP = (ont_adj)? (ont_noun)
PP = NP IN DT? (NP IN DT?){0,5} NP
```

Domain-specific:

```
Space = (spatial_adj)? body_space_or_region
Part = (spatial_adj)? body_part
Term = (Space | Part) IN DT? ( (Space | Part) IN DT? ){0,5} Part
```

Figure 6.4.: Example of the semantic decomposition and recombination process applied to the FMA
NP = Noun phrase; PP = prepositional phrase; IN = preposition; DT = determiner (*a*, *the*, *this* etc); parentheses, ?, | and {n,m} are standard regex operators.

Text to be processed by the recombination patterns is first tokenised and processed by a part of speech (POS) tagger – standard information extraction pipeline components as described in Chapter 5. The recombination patterns are tested by running them against the original list of ontology terms: if the decomposition process and patterns are complete, then every term in the ontology should be matched by the patterns.

6.2.5. Evaluation

The above method was instantiated as a GATE[20] plugin with the recombination patterns expressed in the Java Annotation Patterns Engine (JAPE) language. The JAPE patterns generate annotations in the text: marked ranges each corresponding to an ontology term. In this case, terms matched from FMA morphemes were annotated as `AnatomicalSite`, and terms from DO morphemes as `DiseaseOrSyndrome`. System annotations generated by the semantic decomposition and recombination patterns were compared against manually created `AnatomicalSite` and `DiseaseOrSyndrome` gold standard annotations provided in the 163 progress notes, surgical, radiology and pathology reports in the Ontology Development and Information Extraction (ODIE) corpus[21]. For each annotation type, precision, recall and F_1-measure were calculated against the gold standard annotations (see Chapter 4 for details of these performance measures).

We repeated the evaluation, this time using direct ontology lookup (i.e. using each of the FMA and DO ontologies as very large lookup lists) to create system annotations and stored the results as a separate annotation set for comparison. Finally, system annotations were generated with MetaMap and the results again stored in a separate annotation set (Figure 6.5). As MetaMap has access to the entire UMLS, MetaMap was configured to use only the relevant ontologies and semantic types for the terms being located, so that like-for-like performance could be evaluated against the semantic decomposition/recombination approach. For example, to locate anatomical terms, MetaMap was configured to use only the FMA and UMLS semantic types for anatomical concepts, using the `Anatomy` semantic group assignments from McCray et al.[22]:

```
-Xy -Q 4 -R FMA -J bpoc,bsoj,blor,bdsy,bdsu,tisu,anst,ffas,cell,cellc,emst
```

the -Xy parameters reduce the number of spurious annotations by enabling MetaMap's word sense disambiguation (WSD) server and reduce the number of candidate mappings; the -Q 4 parameter allows MetaMap to identify composite terms (i.e. prepositional phrases) with a maximum length of 4 PPs[3]. The Disease Ontology has been derived

[3] As suggested by http://metamap.nlm.nih.gov/MM11_Usage.shtml

from the UMLS, rather than being one of the terminologies within UMLS. Therefore to identify disease concepts, MetaMap was configured to identify relevant semantic types from UMLS, based on the McCray et al.[22] `Disorders` group assignments:

```
-Xy -Q 4 -J acab,anab,cgab,comd,emod,neop,mobd,dsyn,patf
```

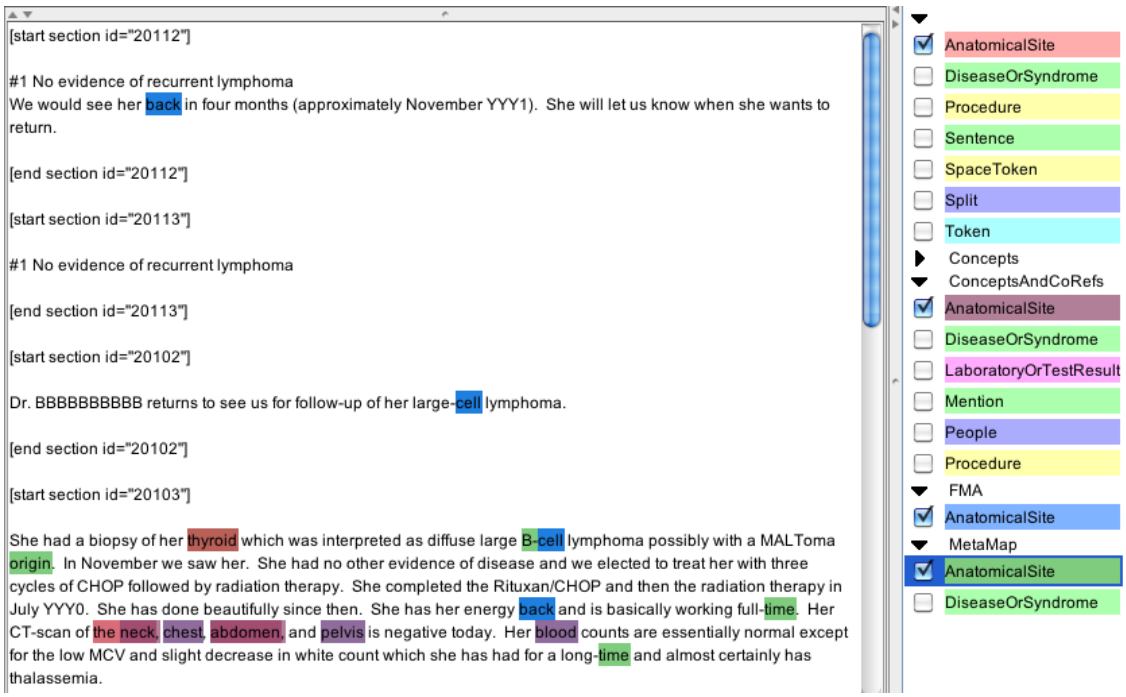


Figure 6.5.: Gold standard vs system annotations for anatomical terms separated into annotation sets for comparison
Top set (unlabeled): System mentions from semantic decomposition/recombination method; ConceptsAndCoRefs: gold standard annotations; FMA: direct lookup annotations from the Foundational Model of Anatomy; MetaMap: annotations produced by MetaMap.

The Wilcoxon signed-rank test was performed over matched document pairs to determine whether there were any significant differences in system performance between the three approaches evaluated.

6.3. Results

6.3.1. Foundational Model of Anatomy

Version 3.2.1 of the complete FMA consists of 150,000 terms made up of 850,000 tokens, of which 4893 are distinct (less than 1% of the total). Semantic decomposition reduced the set of tokens to 1240 morphemes by separating out common prefixes and suffixes as described in Section 6.2.3. The quality assurance step identified 43 spelling errors across 97 terms; these are shown in Table 6.1.

Table 6.1.: Quality assurance step: spelling errors in the FMA

Error	**Correction**	**Example ontology term**	n
arteryy	artery	Left lateral basal segmental pulmonary **arteryy**	1
atery	artery	Deep palmar branch of ulnar **atery**	4
atriovenricular	atrioventricular	Transitional myocyte of **atriovenricular** node	1
bevis	brevis	Trunk of flexor digitorum **bevis** branch of right medial plantar nerve	3
Commisural	Commissural	**Commisural** chorda tendinea of left ventricle	2
Compund	Compound	**Compund** tubuloacinar gland	1
densitiy	density	High **densitiy** lipoprotein	1
diahpysis	diaphysis	Anteromedial surface of **diahpysis** of tibia	1
intermediatel	intermediate	**intermediatel** bronchioles	1
intermideiate	intermediate	Wall of trunk of **intermideiate** atrial branch of right coronary artery	2
laminaof	lamina of	Basal **laminaof** epithelium of bronchus	1
laybrinth	labyrinth	Anterior semicircular duct proper of membranous **laybrinth**	13
leftt	left	**Leftt** middle cerebral arterial trunk	1
luein	lutein	Cytoplasm of **luein** cell	1

Continued on next page

Table 6.1 – continued from previous page

Error	Correction	Example ontology term	*n*
lympahtic	lymphatic	Internodal **lympahtic** vessel	1
Lymphatc	Lymphatic	**Lymphatc** chain at root of inferior pancreaticoduodenal artery	2
medulllary	medullary	Adrenal **medulllary** cell	1
membran	membrane	Left eighth external intercostal **membran**	3
metatearsal	metatarsal	Superficial transverse **metatearsal** ligament	1
middlel	middle	Cavity of **middlel** phalanx of left second toe	2
midlle	middle	Cavity of **midlle** phalanx of left third toe	2
Muscel	Muscle	**Muscel** tissue of crista supraventricularis volume	1
myleocyte	myelocyte	Eosinophilic **myleocyte**	1
myocadium	myocardium	**Myocadium** of apical septal zone of right ventricle	4
nerv	nerve	Plexus branch of anterior branch of left lateral femoral cutaneous nerve with left intermediate femoral cutaneous **nerv**	1
nferior	inferior	Set of **nferior** tributary of tracheobronchial lymphatic vessels	1
ofleft	of left	Dura mater of posterior root **ofleft** fourth sacral nerve	2
oitc	otic	**Oitc** ganglion neuron	2
palpabral	palpebral	**palpabral** vein	3
Penduncular	Peduncular	**Penduncular** tributary of basal vein	3
pumonary	pulmonary	**pumonary** valve anulus	1
qudratus	quadratus	Trunk of **qudratus** femoris part of left inferior gluteal artery	2

Continued on next page

Table 6.1 – continued from previous page

Error	Correction	Example ontology term	n
regon	region	Epithelium of **regon** of epididymis	1
rpoximal	proximal	Cartilage of **rpoximal** phalanx of fourth toe	1
semicular	semicircular	Vein of left **semicular** duct	4
Subdivisionof	Subdivision of	**Subdivisionof** body wall	2
supercilli	supercilii	corrugator **supercilli**	6
Suppresor	Suppressor	**Suppresor** T lymphocyte	1
tissueof	tissue of	Connective **tissueof** serosa of stomach	1
Trunkof	Trunk of	**Trunkof** branch of right vagus nerve to pancreas	1
utricosaccular	utriculosaccular	**utricosaccular** duct	9
venrticle	ventricle	Subdivision of fourth **venrticle**	4
veterbal	vertebral	**veterbal** column	1

Table 6.2 shows the precision, recall and F_1-measure scores, and processing time, for `AnatomicalSite` concept identification by the three system runs against the gold standard annotations, micro-averaged over the 163 documents in the corpus.

Table 6.2.: System performance for identifying `AnatomicalSite` concepts in the ODIE corpus

System	Precision	Recall	F_1	Time(s)
Semantic decomposition/recombination	0.36	0.90	0.51	21
Direct lookup	0.22	0.73	0.34	10
MetaMap	0.30	0.86	0.44	2239

Calculating the two-tailed Wilcoxon signed-rank test (n=163) over matched document pairs for per-document precision, recall and F_1-measure (see accompanying CD for data for each document) showed that the semantic decomposition approach gave significantly better ($p < 0.05$) precision and F_1-measure than MetaMap, although recall was not significantly improved ($p > 0.05$). Against direct ontology lookup, the semantic decomposition

approach gave significantly better precision, recall and F_1-measure ($p < 0.01$). Processing time, however, was slower over the corpus (21s vs. 10s), but was two orders of magnitude faster than MetaMap (21s vs. 2239s).

6.3.2. Disease Ontology

Version 3.1 of the Disease Ontology contains 8610 terms comprised of 26,000 tokens, of which 5055 are distinct (20% of the total). Semantic decomposition reduced this to 769 morphemes. The quality assurance step identified 19 spelling errors across 19 terms; these are shown in Table 6.3.

Table 6.3.: Quality assurance step: spelling errors in the Disease Ontology

Error	Correction	Ontology term
alexithmyia	alexithymia	alexithmyia
ambylopia	amblyopia	disuse ambylopia
anle	angle	primary open anle glaucoma
aquired	acquired	aquired hemangioma
arcinoma	carcinoma	breast arcinoma metastatic to the liver
cogenital	congenital	dominant cogenital severe sensorineural deafness
contricting	constricting	congenital contricting bands
cystadencarcinoma	cystadenocarcinoma	pancreatic colloid cystadencarcinoma
exopthalmos	exophthalmos	endocrine exopthalmos
hemopoetic	hemopoietic	hemopoetic tissue disease
ideopathic	idiopathic	ideopathic interstitial pneumonia
Lffler's	Loeffler's	Loeffler's (or Löffler's) endocarditis
medullblastoma	medulloblastoma	cerebellar medullblastoma
muscoloskeletal	musculoskeletal	muscoloskeletal system benign neoplasm
nephronopthisis	nephronophthisis	nephronopthisis
reproductve	reproductive	female reproductve organ cancer
somatosatinoma	somatostatinoma	jejunal somatosatinoma
trichthiodystrophy	trichothiodystrophy	photosensitive trichthiodystrophy
vericose	varicose	vericose veins

Table 6.4 shows the precision, recall and F_1-measure scores, and processing time, for `DiseaseOrSyndrome` concept identification by the three system runs against the gold standard annotations, micro-averaged over the 163 documents in the corpus.

Two-tailed Wilcoxon signed-rank tests (n=163) over matched document pairs showed that the semantic decomposition approach gave significantly better ($p < 0.01$) precision but

Table 6.4.: System performance for identifying `DiseaseOrSyndrome` concepts in the ODIE corpus

System	Precision	Recall	F_1	Time(s)
Semantic decomposition/recombination	0.58	0.68	0.62	15
Direct lookup	0.69	0.27	0.39	9
MetaMap	0.46	0.83	0.59	1848

significantly worse ($p < 0.01$) recall than MetaMap, although overall showed a significant improvement in F_1-measure ($p < 0.05$). Against direct ontology lookup, the semantic decomposition approach showed significantly worse precision ($p < 0.01$), but significantly better ($p < 0.01$) recall and F_1-measure ($p < 0.01$). As with AnatomicalSite identification, processing time was slower than direct ontology lookup (15s vs. 9s), and again much faster than MetaMap (15s vs. 1848s).

6.4. Error analysis

6.4.1. Anatomical concepts

Examination of the gold standard concepts that were not picked up by the system run with the semantic recombination patterns suggested that the main reasons for false negatives were the annotation of surgical-anatomical concepts such as '*resection margin*', '*stoma*' and '*polyp*' in the gold standard, and the use of abbreviations, all of which were missed by the system.

Table 6.5 gives some additional examples of these false negatives.

Table 6.5.: Example false negatives: terms missed by semantic recombination patterns

Word or noun phrase
PDA
proximal and distal resection margins
SVC
the right MCA distribution
TM's

Precision was nominally low for all system runs (0.22–0.36). Inspection of the results for the semantic recombination pattern system run revealed 1629 false positives, but that

many of these were in fact valid anatomical terms, and only 197 (12%) were actually invalid; many valid terms had not been annotated in the gold standard. For example, in the phrase

> Her CT-scan of the neck, chest, abdomen, and pelvis is negative today

only the terms '*neck*' and '*abdomen*' had been manually annotated, whereas the system also identified '*chest*' and '*pelvis*'. Table 6.6 gives other examples of valid terms picked up by the system. Adding these terms to the gold standard data would have increased the precision to 89% – nearly matching the system recall, which suggests that this method has the potential for providing balanced performance over precision and recall.

Table 6.6.: Nominal false positives that are valid anatomical terms as identified by semantic recombination patterns

Word or noun phrase
angiolymphatic space
dentate line
left nasolabial fold
right posterior eighth rib
right rectus sheath
styloid process of the ulna
the tympanic membranes

Table 6.7 gives examples of noun phrases incorrectly identified by the semantic recombination patterns as `AnatomicalSite` and that would probably be better classified as `Finding` or `PathologicFunction`.

6.4.2. Disease concepts

As with anatomical terms, the use of abbreviations, and inconsistencies and omissions in the gold standard data explained some of the false positives and false negatives, although system precision was generally higher than for the former (0.46–0.69). For example, the annotation of negated concepts was inconsistent: '*she denies [cyanosis]*$_{DiseaseOrSyndrome}$' was annotated, but '*she denies [coronary artery disease]*' was not. Also, some symptoms were incorrectly annotated as a disease concept in the gold standard data, and vice versa. For example '*mood changes*', and '*double vision*' were annotated as `DiseaseOrSyndrome`,

Table 6.7.: Actual false positives: terms incorrectly identified as `AnatomicalSite` by semantic recombination patterns

Word or noun phrase
a haploidentical bone marrow
adenomatous epithelium
diverticula
myopathic process
nonspecific bowel
polypoid
sagittal T1
persistent sequelae
petechiae
pruritic areas

as were '*diagnosis*' and '*side effects*'. The annotation of definite descriptors in the gold standard for demonstrative coreference (see Chapter 8, for example '*her condition*', '*the disease*') also led to reduced system recall as these were not picked up by the semantic recombination rules.

6.5. Discussion

The results suggest that the semantic decomposition/recombination approach to identifying ontology terms in unstructured text provides a significant improvement in both overall accuracy (as measured by F_1-score) and processing time over MetaMap. These findings are similar to those reported for mGrep[14], but with the advantage that the current approach will identify lexical variants, places no restrictions on input text format, and assigns basic semantic type information. Unlike mGrep, we do not store the decomposed ontology in a radix trie, but as plain text as sets of editable regular expressions. However, there is no reason why the underlying regular expression engine used to match text against these expressions should not use a radix trie for efficiency, although here we use the default `java.util.regex` library.

Although not directly comparable, Pyysalo et al.[23] performed direct ontology lookup using the FMA to identify anatomical concepts in a curated corpus of 5000 phrases from PubMed abstracts, achieving a recall of 67%, similar to our findings with the FMA on

the ODIE corpus (73% recall), but significantly lower than our semantic recombination approach (90%). Bashyam[24] used a maximum entropy model with supervised training to identify anatomical phrases in a hand-selected corpus of 4500 sentences, achieving precision and recall 97% – superior to our results, but again they cannot be compared directly as they are on different corpora. The approach presented here requires no supervised training, which has benefits when used in a pipeline with other components for other information extraction tasks (see Chapter 8), and has been shown to generalise for two quite different ontologies.

The method also includes a useful quality assurance step that has found some surprising errors in a mature ontology such as the FMA, which would be difficult to spot by eye, given its size. The errors in the FMA and DO have been verified by the ontology authors and should be absent from future releases (O. Mejino, personal communication 3 Oct 2011; L. Schriml, personal communication 22 May 2012). This is desirable as, in the case of the FMA, the errors are replicated in the 2011 AA release of the UMLS and in other linked data resources that make use of it[1].

In the recombination patterns, the occurrence of an ontology morpheme type is governed by an ontology-independent Kleene operator (?, [n,m]). To allow more fuzzy matches, future work could make use of the positional entropy of each morpheme, i.e. the probability of token t appearing at position p in a given multi-word ontology term, as suggested by Tong et al[17].

6.6. Summary

This chapter has presented a lightweight approach to finding biomedical and clinical concepts in unstructured clinical notes by semantically decomposing ontologies into morphemes, and recombining these morphemes into candidate terms using lexico-syntactic patterns. For concept boundary detection and classification, the approach outperforms MetaMap, the leading open-source biomedical concept recogniser, with the benefit of greatly increased processing speed and much smaller footprint in terms of computing resources.

However, unless they are in the ontology, the method presented here will not identify abbreviations and acronyms. To address this shortcoming, in the following Chapter 7, we apply a similar pattern-based approach to identifying, annotating and classifying biomedical abbreviations, and their expansions, in unstructured text.

7. Identification and expansion of abbreviations in biomedical and clinical narratives

7.1. Introduction[1]

Identification of abbreviations and acronyms, or short forms (SF), for given term definitions, or long forms (LF), is a well researched topic in the biomedical natural language processing domain (see Torii et al.[2] for a review). Gaudan [3] identifies two types of SF usages in biomedical text: *local* short forms, where the defining LF appears with its SF in the document; and *global* short forms, where the SF is used in the document without the defining LF – the reader is assumed to know the meaning of the SF, but this can lead to problems of ambiguity where a given SF has more than one LF in common (or less common) usage.

In general, local abbreviations are typically introduced by giving the LF definition immediately before, or immediately after, the first occurrence of the SF, with either the LF or the SF appearing in parentheses. For example 'bone mineral density (BMD)' or 'EBV (Epstein-Barr virus)': each form an identifiable SF–LF pair. There are a number of existing tools for identifying local abbreviations and extracting dictionaries of SF–LF pairs from them, such as Schwartz & Hearst[4] and Ao & Takagi[5]. Global abbreviations tend to appear in the uninterrupted flow of the text, for example '*A CT scan of the SAS revealed the extent of the bleeding*'; the reader is assumed to know that 'CT' means

[1] This chapter has been published in an abbreviated form as 'BADREX: In situ expansion and coreference of biomedical abbreviations using dynamic regular expressions'[1]

'computed tomography' and 'SAS' 'subarachnoid space' in this context.

It may also be useful to identify a third type of SF usage: *pseudo-global* SFs, whereby both the SF and its LF definition occur in the document, but are used interchangeably and there is no initial pairing of the two on the first appearance of the SF. However, the reader may infer the meaning of the SF by the context and presence of one or more occurrences of the LF. For example:

> The patient is a 63 year-old man with carpal tunnel syndrome ... The patient developed CTS 5 years ago.

A human reader may infer the meaning of 'CTS' from the earlier occurrence of 'carpal tunnel syndrome' without requiring any particular domain knowledge. Computationally, given there may be a number of possible expansions for 'CTS', the correct SF–LF pair could be selected from the dictionary of medical abbreviations based on the occurrence of the specific LF in the text. Alternatively, the first character of each word in every noun phrase could be compared against an index of potential SFs (e.g. strings between 2 and 7 characters consisting of upper-case characters and digits) in the document, to match possible LF expansions. For global SFs, where the LF does not appear at all, approaches based on the verbs and terms surrounding the SF, UMLS semantic type of each LF, and the LF's frequency in the domain at large, can be employed[6][7]. For example, a linguistic pattern (see Chapter 4) that represents the relationship between disorder, experiencer and temporal context might be expressed as

```
[person, disorder_operator, disorder, temporal_concept]
```

where `disorder_operator` is some verb phrase associated with descriptions of a disorder, such as '*develops*', '*suffers from*', '*diagnosed with*' etc. Mapping the second sentence in the example above to this pattern suggests that 'CTS' is a `disorder`-type instance: dictionary entries for 'CTS' short forms not classified in the dictionary as `disorder` types can thus be discarded from the list of candidate SF–LF pairs.

However, tools such as Schwartz & Hearst[4] and Ao & Takagi[5] simply identify and extract SF–LF pairs into a separate dictionary file; they do not classify them according to

their semantic type and they do not provide automatic expansion of short forms within the text *at the point at which they occur in the document*. As such, these methods do not solve the problem of distinguishing local SFs from global or pseudo-global SFs. For example, in a clinical guideline document, 'LAD' might refer to 'Leukocyte adhesion deficiency' (disorder); in a patient's progress notes it might refer to 'left axis deviation' (test result). In the former case, 'LAD' will most likely be a local SF, paired with its LF on its first use; in the latter case, it is likely to be a global or pseudo-global occurrence. There may also be the case where, in the same document, a short form is redefined from its earlier usage. So a method is needed that both identifies SF–LF pairs pairs, resolves unpaired short forms back to their *most recent* definition (if there is one), or, in the case of no definition, resolves a short form to its *most likely* definition from a dictionary.

However, for the purposes of this research, simply resolving a short form to its long form is not enough; we need to know the semantic type of the LF, and subsequently located SFs should inherit this classification. For example, given the text 'WAS (Wiskott-Aldrich Syndrome)', if 'Wiskott-Aldrich Syndrome' has been annotated with the semantic type `DiseaseOrSyndrome`, then 'WAS' should automatically be annotated with this type, as should future 'WAS' mentions. Moreover, given a later occurrence of 'WAS protein (WASP)', where 'WAS protein' has been annotated with the semantic type `AminoAcidPeptideOrProtein`, the 'WAS' should also be given the semantic type `DiseaseOrSyndrome` and be linked back to the earlier mention of 'Wiskott-Aldrich Syndrome' (see Figure 7.2). The knowledge about the disease and the associated protein would then be embedded in later mentions of 'WASP'. Such an approach may facilitate later disambiguation of un-paired abbreviations not possible with dictionary lookup alone[6].

This linkage between a later term and its earlier antecedent, where both point to the same external concept, is known as *coreference* and is dealt with in detail in Chapter 8. Suffice to say here that in situ expansion of a short form, linking it to its fully expanded long form, and to later usages of it, turns out to be an important component in resolving these coreference relationships, particularly in situations of pseudo-global SF usage as

often found in the patient's clinical notes.

In this chapter, a method is developed for identifying, expanding and annotating long-form–short form pairs, and linking subsequent short forms back to their most recent definition in the text. As this chapter is about acronym and abbreviation identification, let us give this method one of its own: BADREX (Biomedical Abbreviation detection with Dynamic Regular Expressions). Here, we do not directly address the problem of disambiguation of global SFs in the general case, as this is a part of a wider area of research on abbreviation word-sense disambiguation (WSD) that is covered in detail elsewhere (e.g. Stevenson et al.[6][7]). However, the semantically typed, in situ local and global SF annotations generated by the method described here are used in Chapter 8, in combination with linguistic patterns as described above, to resolve some of the ambiguities associated with global and pseudo-global SF usage in patients' clinical notes.

7.2. Methods

Throughout this chapter, the notation `<LF, SF>` is used to denote the tuple of the long form and its corresponding short form. BADREX was implemented in Java as a plugin for the General Architecture for Text Engineering (GATE) framework[8]. It takes a `Set` of sentences from GATE's sentence splitter, and for each sentence, five processing steps are performed, where Step 1 is similar to the first stage outlined in Schwartz & Hearst[4] and Step 2 to the third phase of Ao & Takagi[5]:

1. identification of candidate `<LF, SF>` pairs;
2. applying discard conditions to `<LF, SF>` candidates to filter unwanted pairs;
3. identifying the shortest substring in `LF` that best matches SF given the constraints of Steps 1 and 2;
4. matching characters in SF against characters in `LF`;
5. annotating the `SF` and `LF`, storing the `LF` text as feature on `SF` and vice versa, with `SF` inheriting the semantic type of `LF`.

In addition, two further, optional, processing steps can be performed:

1. coreference of unpaired `SF` that match previously found `SF`;
2. expansion of undefined acronynms/abbreviations via dictionary-based lookup

In Step 1, two patterns are created: the 'head' regular expression (regex) identifies a string that contains `{1, maxOuterWords}` words followed by a string of `{1, maxInnerChars}` characters in parentheses or square brackets, and where the first character of the first group is an alphanumeric that matches the first character of the second group. The 'tail' regex consists of a similar pattern but where the first character of the *last word* of the first group is an alphanumeric that matches the *last* character of the second group. For each sentence in the input, if no match is made by the first pattern, then the second pattern may be executed. The 'head' pattern will identify candidate pairs such as:

the behaviour of confluent SV40 transformed rabbit corneal epithelial cells (tRCEC) (1)

and the 'tail' pattern identifies pairs such as:

with two-dimensional proton nuclear magnetic resonance (2D 1H NMR) (2)

(matching characters underlined). In simplified form, the 'head' pattern can be expressed as:

```
\b((\w)\W{0,2}(\w+\W?){1,maxOuterWords})\s*
\((\2.{1,maxInnerChars})(\p{Punct}\s*\w+)?\)
```

and the 'tail' pattern as

```
\b(.{1,maxOuterChars}\b(\w)(\w+\W?))\s*
\((.{1,maxInnerChars}\2(\p{Punct}\s*\w+)?)\)
```

where *maxOuterWords* is the value of the user-defined parameter for the maximum number of words in the long form (default: 10, as per [5]), *maxInnerChars* is the maximum

number of characters in the short form (default: 40, i.e. 10 words), and $maxOuterChars = maxOuterWords \times 4$.

Usually, the short form will appear in parentheses following the long form, but they may appear in reverse order. We allow for this by setting the maximum number of characters as the same by default in both `LF` and `SF`. If the matched short form is longer than the candidate long form text preceding it, the values of `LF` and `SF` are swapped, so that `SF` always points to the abbreviation/acronym, and `LF` always to the definition.

In Step 2 we make use of a simplified subset of the discard conditions for short forms given in Appendix 1 of Ao & Takagi[5]. For example, short forms starting with a preposition, or starting and ending with a digit, are discarded. These conditions are implemented as regular expressions loaded from external configuration files, allowing this behaviour to be easily customised.

In Step 3, dynamically generated regular expressions are used to find the shortest substring of `LF` following a preposition (if present) and where *either* the first character matches the first character of `SF`, *or* the first character of the last word matches the last character of `SF`, depending on whether the 'head' or 'tail' pattern was executed in Step 1. In example (1) above, '*the behaviour of confluent SV40 transformed rabbit corneal epithelial cells*' would be shortened to '*transformed rabbit corneal epithelial cells*'.

In Step 4, non-alpha characters are stripped from `LF` and `SF`, split `LF` into a character array, and iterate over `SF` to match adjacent characters, in the same order, in the `LF` array. If the proportion of matches in relation to the total alpha characters in `SF` $>=$ *threshold* (default: 0.80), then the `<LF, SF>` pair is accepted and added to a `Map` of `<SF, LF>` key/value pairs.

In Step 5, the accepted pair are converted to inline annotations in the text by making use of the `start()` and `end()` methods of the Java `util.regex.Matcher` class, adjusted for term truncation in Step 3 and for the start offset of each sentence. The value of `LF` is stored as a feature on `SF`, and vice versa. If the term definition has already been annotated with one of a configurable set of known annotations, this annotation is used. For example, if '*transformed rabbit corneal epithelial cells*' was previously annotated as

`AnatomicalSite`, then '*tRCEC*' would also be annotated with this semantic type.

In Step 6, code generates regex `Matchers` over the `Map` of pairs populated up to that point, and these are used to locate and annotate unpaired, candidate short forms in sentences forward of the point at which the corresponding long form–short pair was first introduced in the text. The unique identifier of the antecedent `LF` is also stored as a feature on the unpaired `SF`, so that the original definition can easily be located in the text (see Figure 7.1 for an example from a hypertension guideline).

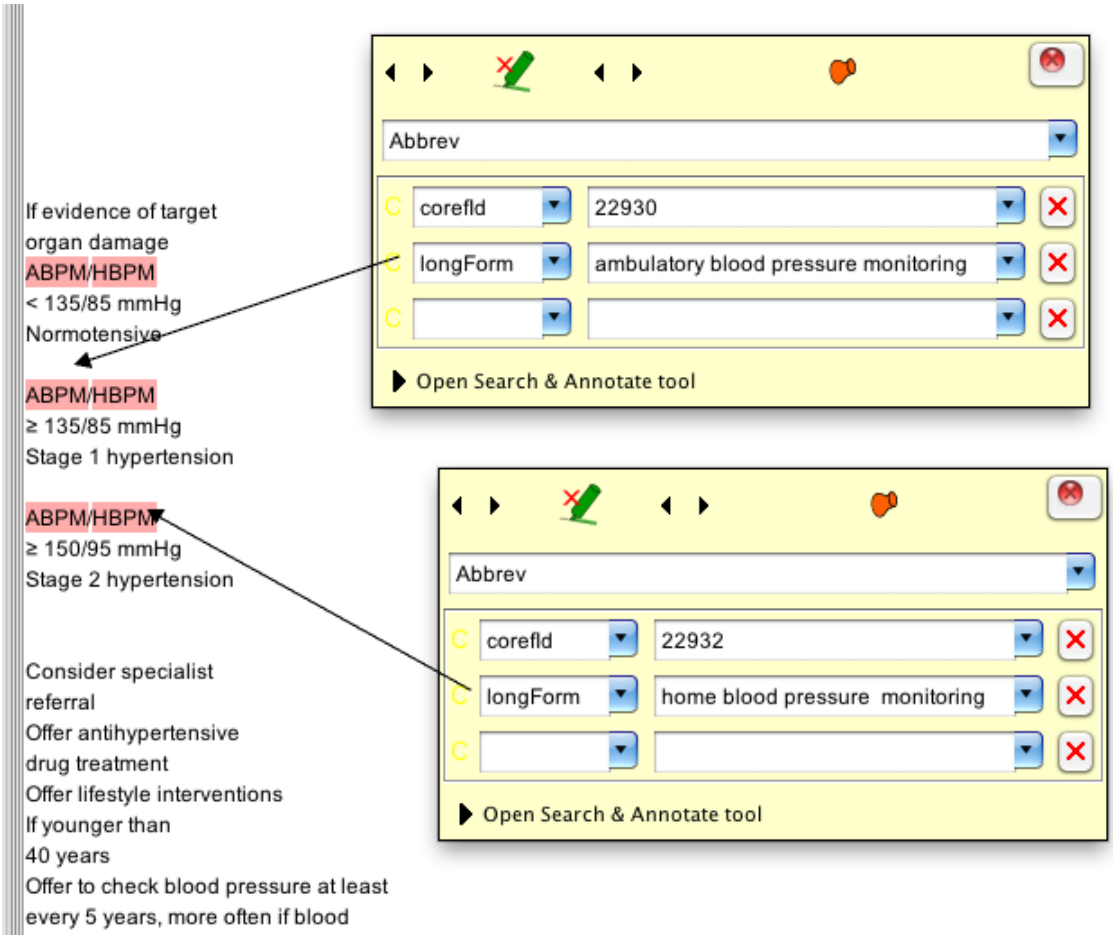


Figure 7.1.: Example of unpaired abbreviations expanded with term definitions and with links back to the location identifiers of each definition's first mention in the text

For the optional Step 7 a dictionary of medical abbreviations is used to expand undefined short forms not matched by the previous steps. To compile this dictionary, medical

abbreviations were extracted from Wikipedia[2], manually grouped according to their semantic type, and each short form stored as a list item with its corresponding long form. For example, 'Pt' or 'pt' would be classified as 'Person' and have a long form of 'patient', 'PT' would be classed as both 'Person' and 'Test' with a long form of 'physiotherapist' and 'prothrombin time':

Person list:

```
...
PCP;term=primary care physician
PMD;term=primary medical doctor
pt;term=patient
Pt;term=patient
PT;term=physiotherapist
...
```

Test list:

```
...
PSA;term=prostate-specific antigen
PSP;term=phenylsulphtalein
PT;term=prothrombin time
...
```

The GATE ANNIE[8] Gazetteer component is used to perform case-sensitive, whole-word-only matches on the document text against this dictionary, creating configurable annotations for each match (e.g. `AnatomicalSite`, `DiseaseOrSyndrome`, `Procedure`, `Test`) and to add the long-form text as a feature on each short form as per Step 5. For short forms with more than one expansion, a number of matches will be made and an annotation will be created for each. As discussed in the Introduction, disambiguating these to identify the correct expansion in the current context is not, however, the function of this module and is performed as a later processing step (see Chapter 8).

[2] http://en.wikipedia.org/wiki/List_of_medical_abbreviations

7.3. Evaluation

Performance of BADREX was evaluated against two well-known gold-standard corpora for testing abbreviation detection algorithm performance: the BioText 'yeast' corpus[4] and the Medstract corpus[9].

The BioText 'yeast' data (`http://biotext.berkeley.edu/data.html`) comprises 1000 MedLine abstracts in a plain text file containing 954 LF–SF pairs annotated with XML-like tags, for example:

```
<Long id=1>endoplasmic reticulum</Long> (<Short id=1>ER</Short>)
```

where the 'id' attribute on the `<Long>` element matches that in the corresponding `<Short>` element. Using a standard XML parser, we identified and corrected errors in malformed 'id' attributes and mismatched or malformed `<Long>` and `<Short>` tags. For example, the XML standard requires all attribute values to be enclosed in double quotes, and that each start tag must be closed by a matching end tag. Correction iterations continued until the file parsed.

The Medstract corpus (http://www.medstract.org/index.php?f=gold-standard) comprises 400 MedLine abstracts in a plain text file, where 414 gold standard LF–SF pairs have been extracted into a separate text file (`http://www.medstract.org/index.php?f=gold-result`: the 'markables'). We analysed the markables file for offset errors, and following correction of these, we compared the abstracts file against the markables to identify any missing pairs.

The precision, recall and F_1-measure performance (see Chapter 4 for details of these methods) of BADREX were evaluated against the corrected BioText and Medstract corpora, and compared the results alongside those for three published systems (two of which were also covered in the review by Torii et al.[2]): Schwartz & Hearst[4] (S-H), ALICE[5] and MBA[10] against the same data. For S-H and ALICE, executable code was available to evaluate on the corrected corpora; for MBA, code was not available so we report the Medstract figures provided by Xu et al.[10]

The effect of varying the `maxOuterWords`, `maxInnerChars`, and `threshold` parameters

against recall, precision and F_1-measure for the larger BioText corpus was also investigated, in order to determine the optimal values for these parameters.

7.4. Results

Figures 7.2 and 7.3 show sample BADREX output in the GATE Developer application for two abstracts from the evaluation corpora. Figure 7.2 shows semantic type assignment (`DiseaseOrSyndrome` and `Protein`) copied from the long form to subsequent short form mentions, and coreference and expansion of unpaired short forms. In coreference mode, short forms occurring within subsequent long forms are also expanded: here, the 'WAS protein' term contains an inner 'WAS' abbreviation that has been expanded to 'Wiskott-Aldrich syndrome'.

Figure 7.3 shows how BADREX allows for whitespace variations in subsequent mentions of the earlier-introduced short form.

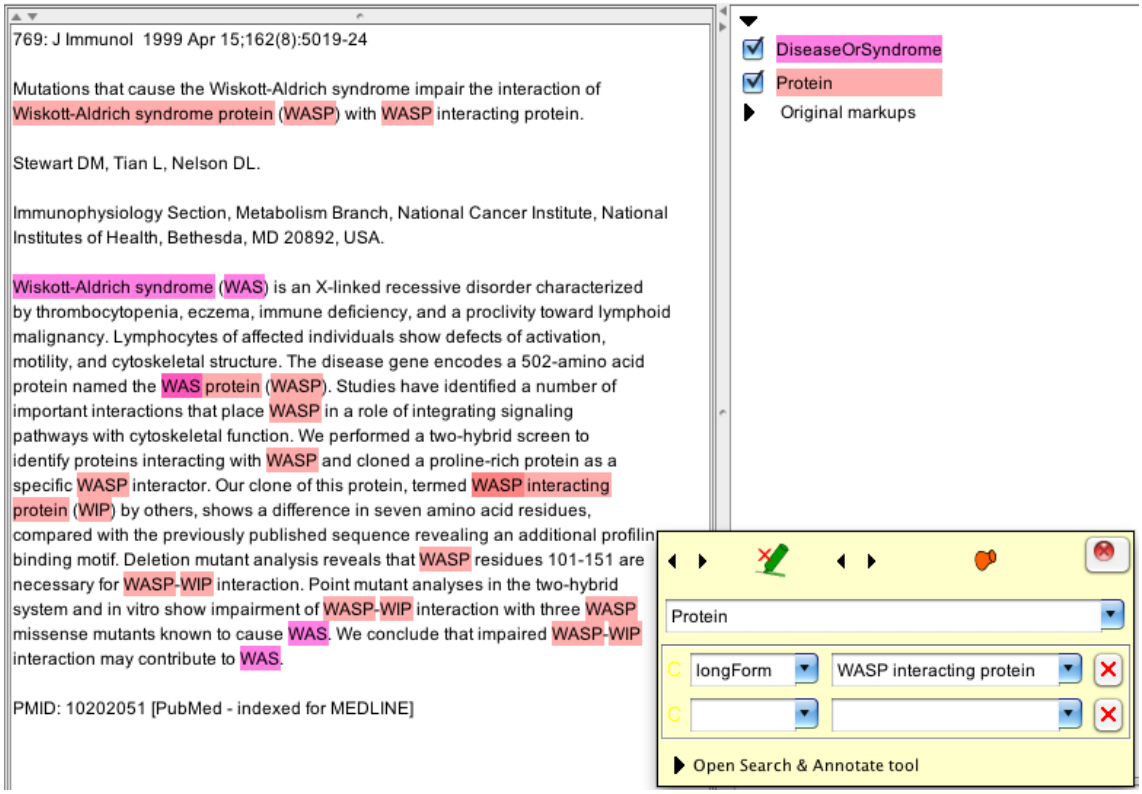


Figure 7.2.: Visualisation of BADREX output in GATE, showing automatically annotated and expanded short forms

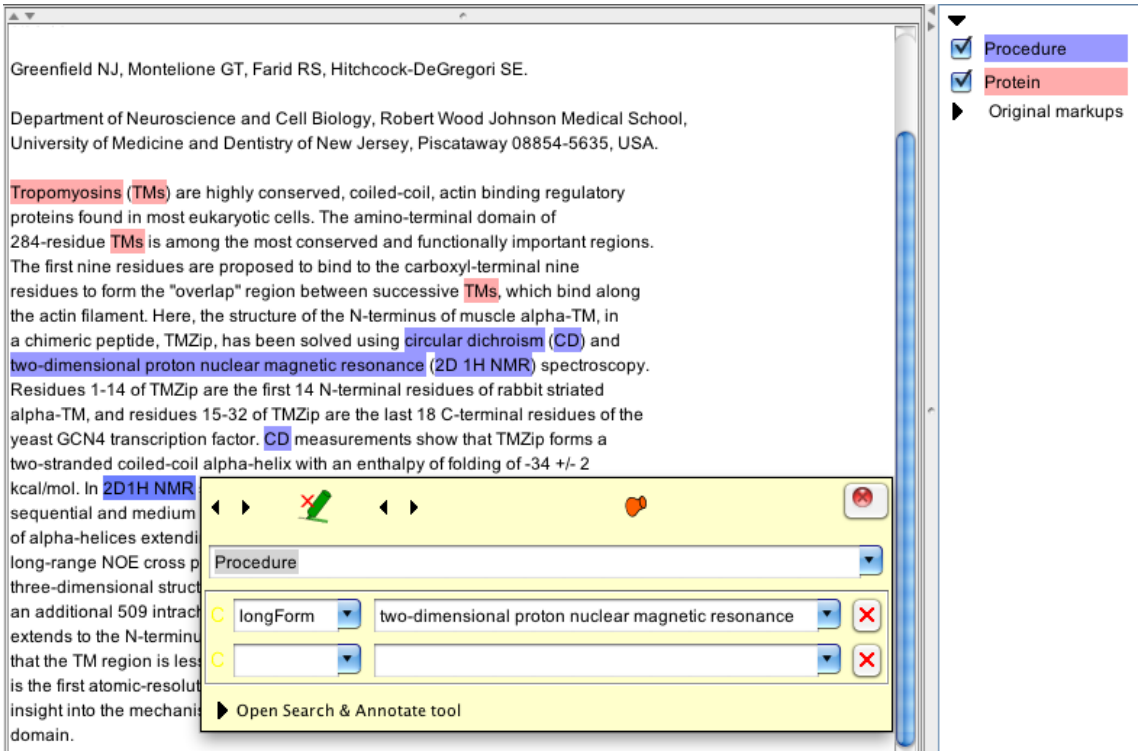


Figure 7.3.: Example showing how BADREX's abbreviation expansion allows for white-space variations in subsequent mentions of the initially introduced short form Here, '2D1H NMR' is coreferenced with '2D 1H NMR' and annotated with the original long form text as a feature

In the BioText corpus, we found 13 incorrectly matching or malformed 'id' attributes and 21 mismatched or malformed `<Long>` and `<Short>` tags. The corrected corpus is available at http://soi.city.ac.uk/ abdy181/software/#badrex

Tables 7.1 and 7.2 shows the results of analysis of the Medstract corpus; against the Medstract gold standard markables an additional 43 markables were identified as correct short-form—long-form pairs, and 17 pairs were amended that were judged to have incorrect spans. The corrected gold standard markables file is available at http://soi.city.ac.uk/ abdy181/software/#badrex

Evaluation of BADREX performance against both corpora in comparison to that of S-H, ALICE and MBA are shown in Table 7.3. As the published MBA results were provided to 2 decimal places, we report all results in this format.

Figure 7.4 shows the effect of varying the maximum number of characters in the short form (`maxInnerChars`) for the default maximum long-form word length (`maxOuter` = 10)

Table 7.1.: Short-form–long-form pairs missing from the original Medstract gold standard markables

Short form	Long form
hCG	human chorionic gonadotrophin
eNOS	endothelial type of NO synthase
3beta-HSD II	3beta-hydroxysteroid dehydrogenase type II
tTGase	tissue transglutaminase
hMG	human menopausal gonadotrophin
IVF ET	in vitro fertilization/embryo transfer
hMG	human menopausal gonadotrophin
hCG	human chorionic gonadotrophin
ds	double-stranded
frag	fragmentation
3-D	3-dimensional
22K hGH	22 kDa growth hormone
alpha-DB	alpha-dystrobrevin
bHLH	basic helix-loop-helix
b FGF	basic fibroblast growth factor
CI	confidence interval
oc	Osteosclerosis
topo II	topoisomerase II
ALP	alkaline phosphatase levels
BMD	bone mineral density
CI	confidence interval
micro-CT	micro-computed tomography
PrE	primitive endoderm
hHb1	Human hair keratin basic 1
bp	base pair
mtDNA	mitochondrial genome
beta 2M	beta 2-microglobulin
pb	peripheral blood
AT	Ataxia teleangiectasia
I.L.S.G.	International Lymphoma Study Group
R.E.A.L.	Revised European-American Classification of Lymphoid Neoplasms'
tHcy	total homocysteine
iNOS	inducible nitric oxide synthase
5-FU	5-fluorouracil
rAAV	recombinant adeno-associated virus
oriP	origin of latent viral DNA replication
HVJ	hemagglutinating virus of Japan
E0'	equilibrium reduction potential
O2-	superoxide
eNOS	endothelial NO synthase
GlOx	glutamate oxidase
beta-END	beta endorphin
tRCEC	transformed rabbit corneal epithelial cells

Table 7.2.: Corrected and original, erroneous long forms in the Medstract gold standard markables

Short form	Corrected long form	Original long form
RAR	RA receptor	regulation of tissue transglutaminase
IAA	indoleacetic acid	in the presence of 10(-6) m 3-indoleacetic acid
EXACCT	exonuclease-amplification coupled capture technique	e exonuclease-amplification coupled capture technique
GlyRalpha2 E3A	glycine alpha2 exon 3A	glycine alpha2 exon 3a (glyralpha2 e3a) and gaba(a) exon gamma
EGFr	EGF receptor	eration through binding to egf receptor
VIN	vulval intraepithelial neoplasia	val intraepithelial neoplasia
SSSS	staphylococcal scalded skin syndrome	scalded skin syndrome
EBER	EBV-encoded small nuclear RNA	ed ebv-encoded small nuclear rna
HD	Hodgkin's disease	15 with Hodgkin's disease (HD †
GluR	glutamate receptor	g chemical selectivity of agonists for the nmda subtype of glutamate receptor
TUNEL	terminal deoxynucleotidyl transferase mediated deoxyu-ridine triphosphate biotin nick end labelling	ted deoxyuridine triphosphate biotin nick end labelling
LC/ESI/MS/MS	HPLC/electrospray ionization tandem mass spectrometric	lective hplc/electrospray ionization tandem mass spectrometric
CYSP	cysteine peptide	conformations of the polypeptides beta endorphin
ESI/MS	electrospray ionization mass spectrometry	ectrospray ionization mass spectrometry
Lid	Lidocaine	lidated for the quantitation of lidocaine
DEX-MPS	dextran-methylprednisolone succinate	DEX-MPS) and its degradation products methylpr †
UV	Ultraviolet	60:40 v/v) and ultraviolet (UV) detection at †

† Incorrect short form

Table 7.3.: Evaluation results against corrected gold standard data sets

System	Corpus	Precision	Recall	F_1
BADREX†	Medstract	0.98	0.97	0.97
	BioText	0.89	0.86	0.88
S-H	Medstract	0.90	0.97	0.93
	BioText	0.91	0.79	0.85
ALICE	Medstract	0.98	0.94	0.96
	BioText	0.92	0.68	0.78
MBA‡	Medstract	0.91	0.88	0.89
	BioText	-	-	-

† Coreference mode disabled in BADREX for this evaluation
‡ Results reported by study authors; software unavailable to evaluate on BioText corpus.

and ratio of characters in the short form that must match, in order, in the long form (`threshold` = 0.8). As shown in the Figure, precision, recall and F_1-measure converge (88, 85, 89% respectively) at around `maxInnerChars` = 40–42 characters. There is no performance benefit in increasing the size of `maxInnerChars` beyond 42 characters, but neither does performance tail off for `maxInnerChars` values beyond this point.

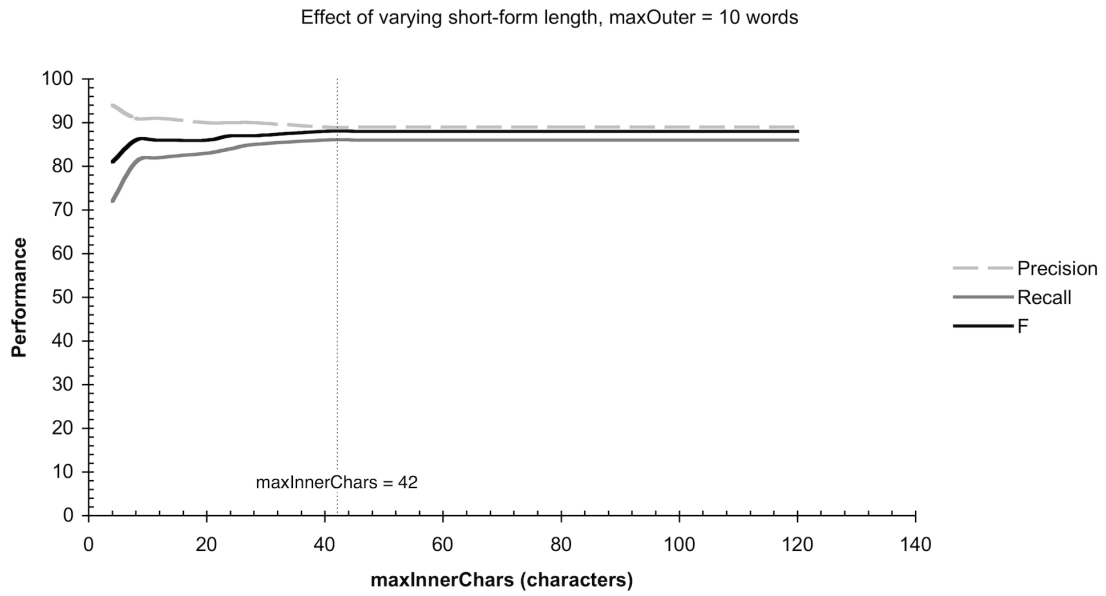


Figure 7.4.: Effect of varying short-form length for constant long-form length and threshold

Figure 7.5 shows the effect of varying the maximum number of words in the long form (`maxOuter`) for the default maximum short-form word length (`maxInnerChars` = 40) and

short-form match threshold (`threshold` = 0.8). As shown in the Figure, precision, recall and F_1-measure are maximal (90, 88, 86% respectively) at around `maxOuter` = 7 words. In a similar distribution to that shown in Figure 7.4, there is no performance benefit nor penalty in increasing the size of `maxOuter` beyond 20 words.

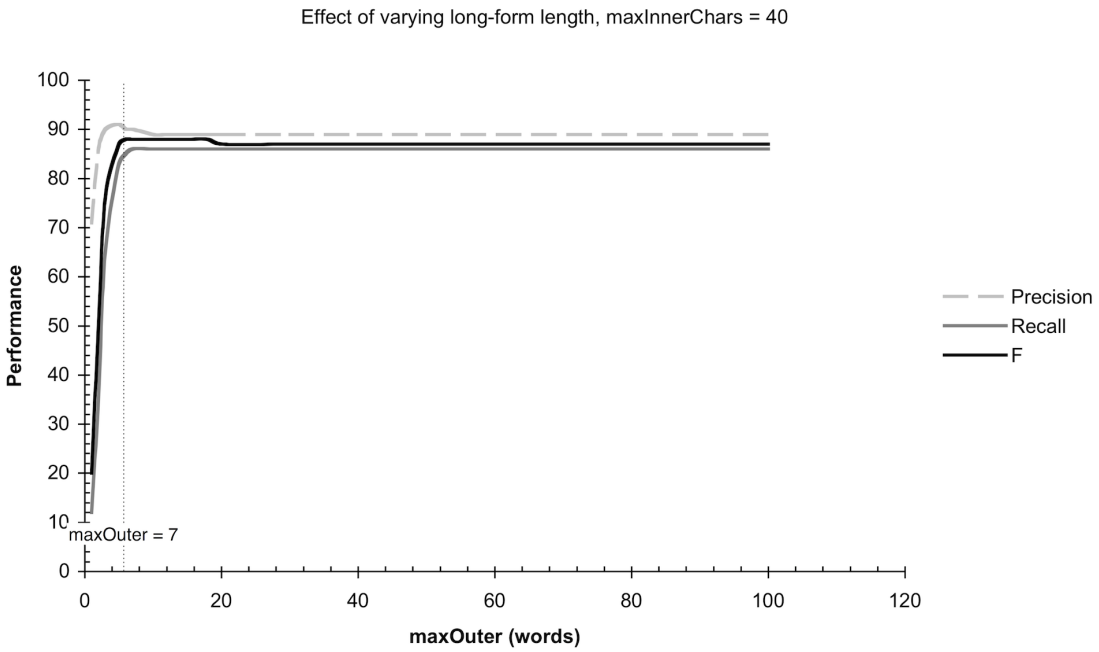


Figure 7.5.: Effect of varying long-form length for constant short-form length and threshold

Figure 7.6 shows a similar plot to Figure 7.4 but using the provisionally optimised `maxOuter` parameter of 7 words and constant `threshold` of 0.8, giving slightly better precision, recall and F_1-measure of 91, 86 and 88% respectively at `maxInnerChars` = 42 characters.

Figure 7.7 shows the effect of varying `threshold` for near-optimal `maxInnerChars` and `maxOuter` values of 40 and 7 respectively. As shown in the figure, precision, recall and F_1-measure converge at around 87% giving an optimal `threshold` of 0.65. However, beyond this point, precision rises a little more steeply than recall falls, so it may be better to use a slightly higher threshold than 0.65 in order to give preference to higher precision. At `threshold` = 0.76, precision, recall and F_1-measure are 90, 86 and 88%, which confirms this.

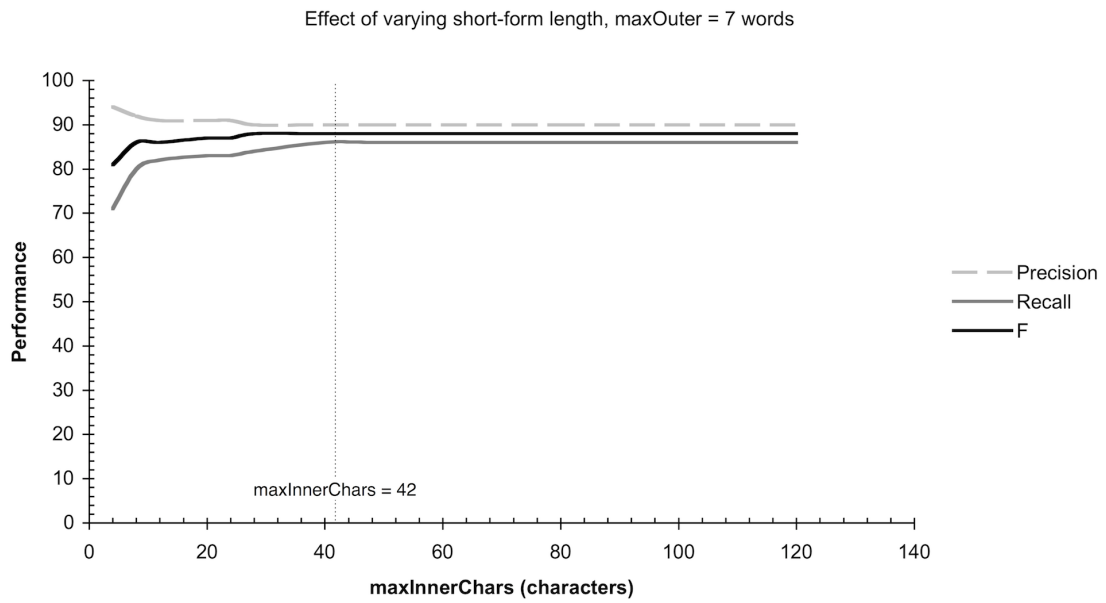


Figure 7.6.: Effect of varying short-form length for maximum 7 word long forms

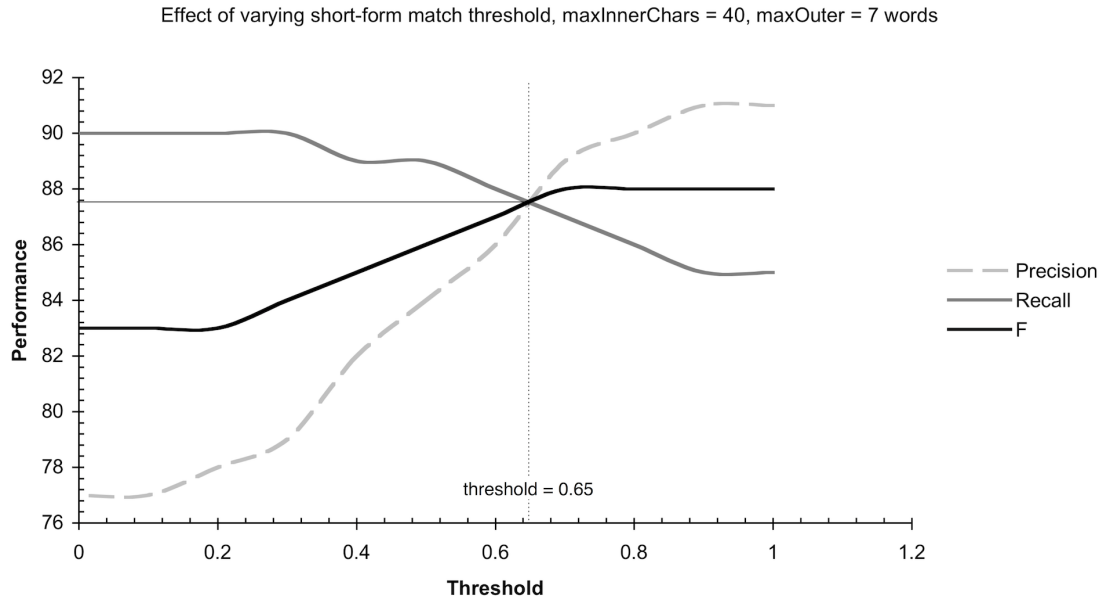


Figure 7.7.: Effect of varying short-form match threshold for constant short-form and long-form length

7.5. Discussion

The goal of this work was to develop a customisable tool for identifying, expanding and annotating in situ biomedical abbreviations in free text, while matching or exceeding the performance of existing approaches. As show in Table 7.3, performance on both of the evaluation corpora is slightly better than the next two best-performing approaches. The MBA tool was not available for direct evaluation against the corrected corpora: had this tool been available, it may have shown improved performance than the figures quoted by the MBA authors against the original corpora.

Running both 'head' and 'tail' candidate matches, allowing a variable threshold and only considering alpha characters when matching allows long form–short form pairs such as '*topoisomerase I (Top1p)*', '*Phosphatidyl-inositol-3-kinase (PI3K)*' and '*two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)*' that are missed by the other approaches, yet without unduly compromising precision.

Steps 1 and 3 of the approach presented here are similar to the algorithm described in Schwartz & Hearst[4], which they implemented as a finite-state machine (FSM) with the constraints hard-coded. However, by using dynamic regular expressions in these steps, we simplify the creation of the FSM and allow it to be easily parameterised, so that optimal parameter values for a given corpus can be identified. In this study, default values were used for these in the comparative evaluation, but by plotting different values for the parameters, we found that, for the BioText corpus, precision, recall and F_1-measure could be optimised with maximum candidate long-form (`maxOuterWords`), short-form (`maxInnerChars`) and character-match threshold (`threshold`) values of 7 words, 42 characters, and 0.65–0.76 respectively. In general, iterative regression techniques could be used to find optimal values for these parameters to maximise precision and/or recall for other corpora.

However, the use of regular expressions, although both simplifying the implementation and making it more flexible, does have some limitations. For example, the regexes used here do not currently allow for nested parentheses occurring either in the short form or long form. Also, the current implementation requires that either the short form or long form appear in parentheses; pairings where the two are separated by other punctuation,

Table 7.4.: Example pairings missed by BADREX on the BioText corpus

Pairs
protein phosphatase (PP1(C))
Temperature-sensitive (Ts(-))
mitochondrial actin binding protein(s) (mABP)
pBEVY (bi-directional expression vectors for yeast)
(Suppressors of Arf ts, SAT)
phosducin-like orphan proteins (PhLOP1 and PhLOP2)
(phox) (PX)
p59fyn (Fyn)
chloramphenicol acetyl transferase (CATIII)
poly(A) polymerase (PAP)
upstream activating sequence (UAS(GAL1/10))
TASR (TLS-associated serine-arginine protein)
C-terminal domain (CalphaB)
phosphatidylinositol 3,4,5-trisphosphate (PtdIns(3,4,5)P3)
OCT (22-oxa-1alpha,25-dihydroxyvitamin D3)
Ycf1p (yeast cadmium factor or glutathione S-conjugate pump)
Na(+)/H(+) exchanger regulatory factor 2 (NHERF2)
eukaryotic translation initiation factor 2B (eIF2Bepsilon)

such as a comma, will be missed. While both precision and recall were very high on the Medstract corpus (0.98 and 0.97 respectively), recall in particular was significantly lower on the larger BioText corpus (0.86). The above-noted limitations in the current regex implementation was one of the reasons for this reduced recall: examples of pairings missed by BADREX on the BioText corpus are shown in Table 7.4. Some of these missed pairings will have been picked up by increasing the values of the `maxOuterWords` and `maxInnerChars` parameters (e.g. for '*phosducin-like orphan proteins (PhLOP1 and PhLOP2)*'. Reducing the `threshold` parameter would have picked up some other missed pairs (e.g. '*pBEVY (bi-directional expression vectors for yeast)*' but this would have been at the expense of reduced precision elsewhere, as suggested by Figure 7.7. Other causes of error were spelt-out Greek letters appearing in the short form (e.g. '*eIF2Bepsilon*') and capitalised roman numerals (e.g. '*CATIII*').

7.6. Summary

In this chapter, we have demonstrated an approach to identifying term definition–term abbreviation pairs that uses regular expressions dynamically generated from document content. This approach yields a modest performance improvement in comparison to previous approaches. The main benefit, however, is that the code provides in-place annotation, expansion and coreference in a single processing pass through each document. In addition, this approach requires no training data; however, via runtime customisation of its input parameters it can be trained if required so that optimal parameter values can be estimated for different corpora.

As a result of the research presented in this chapter, a substantial number of errors in two reference corpora, which have been used to evaluated the performance of a number published methods for identifying biomedical abbreviations, have been corrected. These corrected corpora are now available to researchers for use in future evaluation studies.

We have not yet evaluated the coreferencing features of BADREX, nor the annotation of common medical abbreviations extracted from Wikipedia. Future work will need to evaluate the contribution of these features as components in the disambiguation of undefined abbreviations, such as typically encountered in a patient's clinical notes. In Chapter 8, we make use of BADREX as a component in a general coreference resolution system for clinical narratives in which these features are utilised.

8. Putting it all together: coreference resolution for identifying processes of care and chains of events in clinical narratives

8.1. Introduction[1]

With the work presented in the previous chapters, we now have a number of the components for performing a range of information extraction tasks in the clinical domain, such as:

- optimal text segmentation for pre-processing input to MetaMap
- identification of quantitative and temporal concepts
- identification and correction of spelling errors
- identification of anatomical and disease concepts
- expansion of biomedical and clinical abbreviations

In this chapter, we discuss contextual features, knowledge resources and lexical patterns for resolving coreference relations in the clinical domain, making use of all the above components. We apply the clinical knowledge extraction framework to identify coreference relations in a wide variety of clinical reports, and argue that this is an important component in addressing the wider problem of identifying processes of care in the clinical narrative.

[1]This chapter has been published in an abbreviated form as 'Lexical patterns, features and knowledge resources for coreference resolution in clinical notes'[1]

In linguistics, the relationship of coreference holds when two or more expressions or *mentions* (typically noun phrases) refer to the same external entity, independent of the order of the expressions within the text. The semantic relation between the expressions is one of identity. This identity relation leads to the assumptions of symmetry (if A is coreferent with B, then B is coreferent with A) and transitivity (if A is coreferent with B, and B is coreferent with C, then A is coreferent with C). Coreference can be considered a specific type of *anaphoric relation* where a later expression (anaphor) has some semantic relation to an earlier expression (antecedent) and disambiguation of the anaphor is dependent on knowledge of the antecedent[2][3]. In a general anaphoric association, the semantic relation may be of identity, but not necessarily; for example, anaphor and antecedent may be in a part–whole relationship.

Pronominal coreference considers the resolution of pronouns back to their correct antecedent (e.g. resolving '*they*' to '*the clinical team*'), while bridging coreference considers the resolution of definite descriptors (e.g. '*the procedure*') and semantically equivalent terms (e.g. synonyms and hypernyms) back to the specific antecedent. Resolution of bridging coreference may make use of the following features:

- *lexical features*, such as matching headwords where one term is preceded by one or more modifiers and the other is unmodified (e.g. resolving '*the swelling*' to an earlier mention of '*the lower extremity swelling*');
- *external domain knowledge*, such as hypernym relations in resolving '*the antibiotic*' to an earlier mention of '*amoxicillin*' and '*the infection*' to an earlier mention of '*staph bacteremia*';
- *synonym relations*, such as resolving '*shortness of breath*' to '*dyspnea*'.

Relations may be both coreferent and anaphoric, for example '*initially the patient refused* ***bronchoscopy*** *but agreed to* ***it*** *later*', as the anaphor '*it*' can only be understood in relation to the antecedent '*bronchoscopy*', and both '*it*' and '*bronchoscopy*' refer to the same, external concept (a bronchoscopy procedure). With compound terms, the relationships can be multiple and more complex, for example:

> **the patient** was admitted with a *head* wound laceration ... we sutured **her** *scalp* laceration

where there is potentially both an anaphoric and coreferent relationship between '*her*' and '*the patient*' (given the world knowledge that the patient is female), '*scalp*' is anaphoric to '*head*' in a meronym—holonym (part—whole) relationship, and '*head wound laceration*' and '*scalp laceration*' are potentially coreferent if they refer to the same injury.

The rule-based heuristics that characterised initial approaches to coreference resolution have largely been supplanted by a variety of supervised machine learning approaches. In general, rule-based approaches have been dominated by research on pronominal coreference on general texts by Lappin, Leass and Mitkov (reviewed in [2]), which typically involve a backward-looking search from a given pronoun to the best antecedent. Antecedent ranking rules consider factors such as gender, number, token distance and sentence recency; syntax such as grammatical role or dependency relation (subject, direct object, indirect object), person, and position; and discourse models such as centering theory[4] (see Section 8.3.4).

Supervised machine learning approaches have, until recently, been dominated by the *mention-pair* model, which treats coreference resolution as a binary classification problem between pairs of mentions (as in the example above), but has been criticised for considering each mention in isolation and not the wider contexts in which each mention occurs[5]. More recent models consider the task as a semi-supervised, cluster-ranking problem in which mentions are grouped into clusters, and then a ranking algorithm, which considers grammatical, syntactic, semantic and discourse-based contextual features, is used to classify clusters into those that are coreferential and those that are not[4][5]. However, despite the move toward machine-learning approaches, the Stanford NLP Group's state-of-the-art system for resolving coreference relationships in the general domain is entirely rule-based[6].

Resolution of '*it*', '*this*' and '*that*' pronouns present particular problems for coreference resolution, as they are often used redundantly and may not refer to any specific mention – i.e. they may be used in a *pleonastic* sense. For example, in the phrase '*It is important to note that thresholds vary*', both '*it*' and '*that*' are pleonastic and do not refer back to

an earlier, specific mention of a concept. In practice, distinguishing pleonastic references from anaphoric references is not straightforward. Consider the following, albeit somewhat contrived, examples:

> Patient is taking vancomycin. It has been prescribed to treat the MRSA infection.

> Patient is taking vancomycin. It has proven difficult to treat the MRSA infection.

In the first example, '*it*' is anaphoric to '*vancomycin*' but in the second, '*it*' is pleonastic, despite both sentences having very similar structure.

8.2. Relevance to the clinical domain

Resolution of coreference is particularly important in clinical narratives, such as progress notes and discharge summaries, as it is often required to uncover implicit and contextual information. For example:

> S. Holmes, a 53 year-old male, was seen on 23/06/2012. The patient suffers from chronic lower back pain. He has been taking Vicoprofen for this since 07/2011, but the medication is not managing his discomfort.

To the human reader, it is clear that the patient's lower back pain is managed unsuccessfully with Vicoprofen, and both the pain and the Vicoprofen prescription predates his most recent visit. To infer this information computationally, the following steps need to be performed:

1. Identify and classify clinical concepts and temporal expressions in the statement(s), (as described in Chapters 5 and 6)
2. Identify terms with the same classification that are coreferential
3. Identify temporal relations between terms

For the first step, carrying out part-of-speech tagging, verb-group (VG) chunking and concept identification, as described in Chapter 5, would identify the following annotations:

> [S. Holmes]$_{\text{Person}}$, a [[53 year-old]$_{\text{Age}}$ male]$_{\text{Person}}$, [was seen]$_{\text{VG}}$ on [23/06/2012]$_{\text{Date}}$.
>
> [The patient]$_{\text{Person}}$ [suffers]$_{\text{VG}}$ from [chronic lower back pain]$_{\text{Problem}}$.
>
> [He]$_{\text{Pronoun}}$ [has been taking]$_{\text{VG}}$ [Vicoprofen]$_{\text{Treatment}}$ for [this]$_{\text{Pronoun}}$ since [07/2011]$_{\text{Date}}$, but [the medication]$_{\text{Treatment}}$ [is not managing]$_{\text{VG}}$ [[his]$_{\text{Pronoun}}$ discomfort]$_{\text{Problem}}$

However, without a method for resolving '*he*', '*his*' back to '*the patient*' (and '*the patient*' back to '*S. Holmes*' and '*a 53 year-old male*') , '*this*' and '*his discomfort*' back to '*lower back pain*', and '*the medication*' back to '*Vicoprofen*', there is no way that the above reasoning can be inferred computationally. Moreover, this example shows the importance of resolution of complete *chains* of coreference ('*S. Holmes—53 year-old male—The patient—he—his*', '*chronic lower back pain—this—his discomfort*', '*Vicoprofen—the medication*') to enable this information to be extracted. The relationships between the coreference chains can be visualised in the graph shown in Figure 8.1. From this, and using the narrative schema notation developed by Chambers & Jurafsky[7] for modelling chains of events, a temporally ordered set of events and roles can be described as shown in Table 8.1.

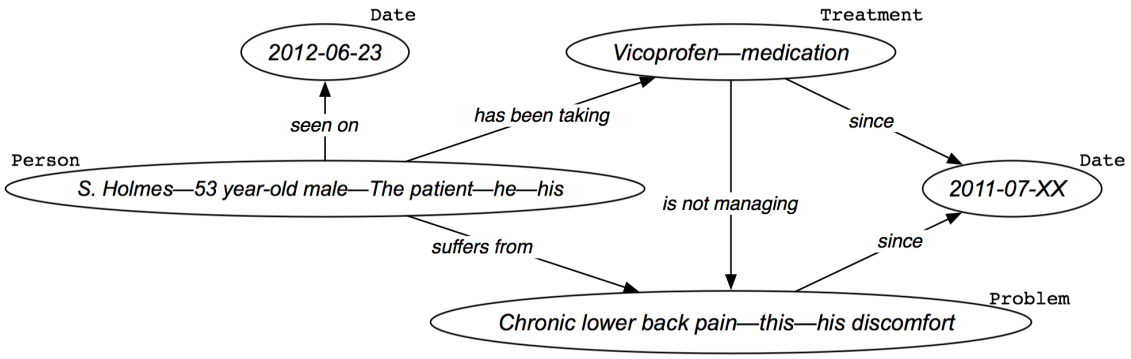


Figure 8.1.: Graph-based representation of relationships between coreferential chains

In a review of coreference methodologies, Zheng et al.[8] noted that there was a lack of both manually annotated corpora and automated systems for identifying coreference within the clinical domain. They concluded that an approach that identifies patterns specific to clinical texts, combined with adaptation of more general methods, would be

Table 8.1.: Narrative event chains from the graph of coreference and relations in Fig. 8.1

Date	Events	Roles
?	S suffers from P	S = Holmes, 53 year-old male, the patient
	↓	P = chronic lower back pain
	↓	
2011-07-xx	S has been taking T	T = Vicoprofen
	↓	
?	T is not managing P	
	↓	
2012-06-23	P was seen on	

a necessary first step towards a solution. However, existing, general-purpose coreference tools, such as the BART Coreference Toolkit[9] or the Stanford Deterministic Coreference Resolution System[6] — even when retrained for the clinical domain — perform poorly on clinical texts, where recall is particularly low, varying from 0% to 35% [10]. This is perhaps not surprising, as transcribed clinical notes present particular problems for identification of co-referring terms, such as:

- The domain knowledge requirement, to identify synonyms where there is no overlap between mention strings; for example, '*the patient's abdomen*' and '*the patient's epigastric area*'.
- Exactly matching strings might not corefer: the contexts of the events or clinical conditions they represent may may be different or affect different experiencers. For example '*hypertension*' in 'There is a family history of [hypertension]$_{\text{Problem}}$' and 'the patient's [hypertension]$_{\text{Problem}}$ is being managed with ACE inhibitors'.
- The potential for spelling inconsistencies and errors.
- The use of ambiguous abbreviations without prior definition. For example 'Pt' may abbreviate '*patient*', but 'PT' may abbreviate '*physiotherapy*' or '*prothrombin time*'.
- Name anonymisation potentially resulting in the same personal name being replaced with a different string during de-identification, and the anonymised name may not match the patient's gender.

- The potentially wide scope of resolution for personal pronouns. For example, '*he*' might refer to '*the patient*' mentioned several sentences or paragraphs previously, as intervening paragraphs may have discussed, for example, laboratory results.

Until recently there have been few evaluation reports of automated approaches to coreference resolution in clinical texts. Romauch[11] developed a knowledge-based system using the MetaMap Transfer (MMTx) application and the Unified Medical Language System (UMLS) to resolve definite descriptors in clinical practice guidelines, reporting an F_1-measure of 75.8%. Error analysis revealed inadequate acronym/abbreviation detection leading to incorrect UMLS mappings made by MMTx (such as abbreviations for clinical terms being incorrectly identified as gene names) and coreference of terms sharing the same hypernym (e.g. '*further surgery*' and '*incomplete excision*' – although the error here is more related to linking a possible, future planned event with a previous event) and incomplete coreference chains as sources of error. For hospital discharge summaries, He[12] used a supervised decision-tree classifier with a mention-pair model to resolve coreference chains of `Person`, `Symptom`, `Disease`, `Medication`, and `Test` mentions, and achieved a mean F-measure of 81.0% (ranging from 95.0% for Medications to 50.6% for Tests). Analysis revealed incomplete handling of temporal context, lack of knowledge-based handling of synonym and hypernym relationships, and lack of acronym/abbreviation detection, as the main factors affecting system recall.

To help address the lack of research into coreference resolution in clinical texts, two manually annotated corpora of clinical anaphoric relations have recently been made available[13]: the Ontology Development and Information Extraction (ODIE) corpus[14] and the i2b2/VA corpus[15]. The former consists of de-identified clinical notes and pathology reports from the Mayo Clinic, and discharge summaries, progress notes, radiology reports, surgical pathology reports, and progress notes from the University of Pittsburgh Medical Center (UPMC). The latter consists of de-identified discharge summaries from Partners HealthCare, Beth Israel Deaconess Medical Center, and UPMC.

The corpora have been divided by Uzuner et al.[13] into a training set of 589 documents previously annotated for mentions of persons and clinical concepts (`Procedure`,

`DiseaseOrSyndrome`, `SignOrSymptom`, `Reagent`, `LaboratoryOrTestResult`, `OrganOrTissueFunction`, and `AnatomicalSite` in the ODIE corpus; `Problem`, `Treatment` and `Test` in the i2b2/VA corpus) and a test set of 388 documents. For the training set, adjudicated 'ground truth' coreference chains have been provided in the following format, in which the concept's class, and the text string and line/word offset of each coreferent markable are identified, for example:

```
c="right hip osteoarthritis" 21:0 21:2||c="advanced osteoarthritis
of his right hip" 49:3 49:8||c="severe osteoarthritis of the right
hip" 51:6 51:11||t="coref Problem"


c="the patient" 22:0 22:1||c="she" 23:0 23:0||c="she" 24:0 24:0||
c="her" 26:0 26:0||c="she" 27:0 27:0||c="she" 29:0 29:0||t="coref Person"
```

An evaluation script has also been released[16] which compares system output in the above format to a given reference set of test output. The script measures system performance against a number of metrics well-known in the general field of coreference resolution, including B^3[17], MUC[18], CEAF[19] and BLANC[20]. Using these metrics, Zheng et al.[21] have recently published results for a system that used a variety of supervised machine learning approaches to resolve coreference in the ODIE corpus. Using a support vector machine with a radial basis function, they achieved a mean F_1-measure (over all metrics) of 53.1%.

Briefly, these metrics perform complex set-wise comparisons of coreference chains between the key set and the system output under evaluation. Given that the coreference relation is one of identity, as noted in Section 8.1, the relation between mentions in a coreference chain must be transitive and symmetric - i.e. if $\rightarrow$ denotes the coreference relation, and A $\rightarrow$ B and B $\rightarrow$ C then A $\rightarrow$ C and C $\rightarrow$ A. Thus, for the purposes of evaluation, coreferences chains are treated as unordered sets. So, in a coreference chain of (A, B, C) in the key (ground truth) set, a system that generates the chain (C, A) should be gain a partial recall score but should not be penalised on precision. Furthermore, a system that generates (A, B, E, C) where E is not in the key set should gain a partial precision score

but should not be penalised on recall. On the other hand, given a key set of (A, B, C, D), a system that generates split chains of (A, B) and (C, D) might be partially scored for both recall and precision. Each of the metrics currently in use different algorithms to attempt to weight these types of scenarios when scoring precision and recall (for a detailed discussion of how each metric calculates these weights, refer to Cai and Strube[22], and Zheng et al.[8]). In in doing so, some metrics can give anomalous results in certain edge cases, as noted below.

There is controversy over which is the most valid metric for evaluating system performance, particularly when dealing with coreference of system-generated mentions not in the key set, leniency in handling split coreference chains, and singletons (mentions with no coreferents). For example, with the MUC metric, Luo [23] noted that adding all mentions into a single coreference chain resulted in a recall of 100% and precision of 79% giving an F_1-measure of 88%. Unexpected scores are also given for null system output (i.e. no coreference relations and no mentions): against the ODIE ground truth set containing 44,000 mentions and 5200 coreference chains, F_1-measures of 0.936, 0.5, and 0.686 are reported by B^3, Blanc, and CEAF, respectively, for null system output, whereas a score of 0 in each might reasonably be expected. As a result of these anomalies, some researchers, e.g. Poon et al.[24], have instead reported pairwise evaluation scores, i.e. evaluation of correctly marked mention pairs, rather than complete chains, thus ignoring transitivity.

In the following sections, the components of the framework that identify coreference relations between the types of clinical concepts previously discussed in Chapters 5 and 6 are presented.

8.3. Methods

8.3.1. Preliminary analysis of the training corpora

Given the particular problems of pronominal coreference resolution, as discussed above, preliminary analysis was carried out on the distribution of pronouns across the complete set of documents in the training corpora. Additionally, analysis of the distribution of

Table 8.2.: Distribution of `Person` and pronoun mentions in the training data ground truth mentions and coreference chains

	All mentions	Person[a]	Personal pronouns	Patient[a]	Patient personal pronouns	Other pronouns
Coreference chains	35525	19484	10313	14580	8879	1585
Mentions	70623	21127	10421	n.d.	n.d.	3862

[a] All mentions — includes personal pronouns. n.d. = not done - data not available

references to the patient, both directly and via personal pronouns, was performed. Using regular expressions, coreference chains of Person-type mentions in the training data containing the words '*patient*' or '*pt*' were classified as 'patient' references and these were extracted and quantified. Similarly, personal pronouns across both all generic `Person`-type coreference chains and those within 'patient'-type coreference chains were extracted. Table 8.2 shows the distribution of 'patient' and person pronoun references across both corpora in the training set. Figure 8.2 shows the overall distribution of all pronouns in the training set.

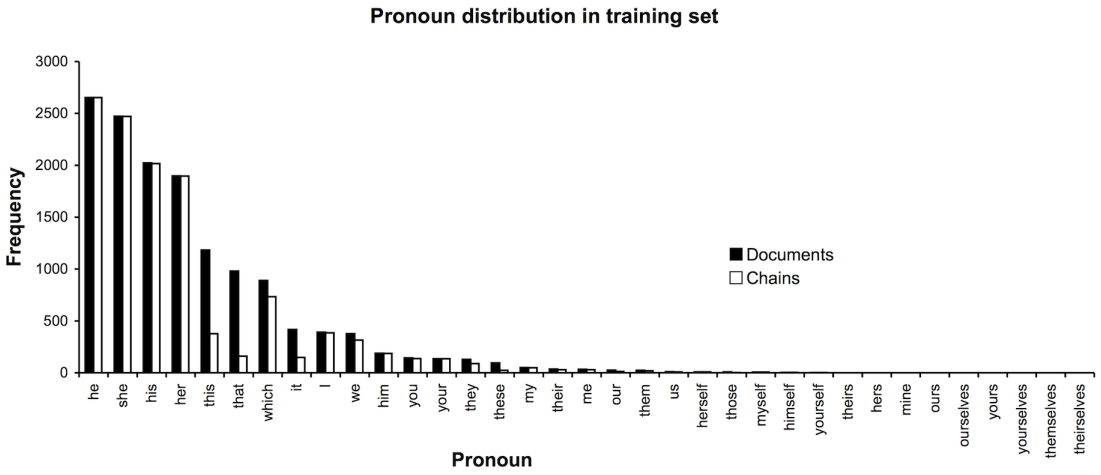


Figure 8.2.: Distribution of pronouns in the training set

As shown in Table 8.2, in the training set, of 35525 coreferenced mentions, 19484 (55%) refer to a `Person`, and 8879 out of 10313 (86.1%) personal pronoun mentions that appeared

in a coreference chain refer to the patient. Of all coreferenced `Person mentions`, 14580 out of 19484 (74.8%) are in a 'patient' coreference chain. Examination of the remaining mentions revealed that they referred to members of the clinical team, to family/significant others, or to the person receiving the report.

As shown in Figure 8.2, pronoun usage is dominated by third-person singular, nominative or possessive pronouns (*he*, *she*, *his*, *her*), all of which participate in coreference, as shown by the equal height of the black bars (overall document frequency) and white bars (coreference chain frequency). Also notable is the discrepancy between the overall occurrence of '*it*', '*this*' and '*that*' and their low occurrence in coreference chains, suggesting that their usage in these corpora was predominantly in a pleonastic sense (see Section 8.1).

8.3.2. Architecture overview

The basis of the clinical coreference component of the framework is, as before, a rule-based pipeline that runs within GATE[25]. Rules were developed using the Java Annotation Patterns Engine (JAPE) language, and external domain knowledge integration plugins using Java. JAPE allows pattern matching and evaluation of text annotations using a regular expression-like syntax. An annotation represents a marked range in the text, corresponding to some entity or mention, with start and end nodes, a document-unique identifier, and a set of features (attributes on the annotation). Each node points to a character offset in the document. One of the benefits of JAPE is that annotations not specified in the input are ignored for pattern matching purposes, which enables patterns to be generalised when, for example, intervening punctuation and prepositions are not significant. The patterns provided in the following sections exemplify this generalisation.

As noted in Section 8.2, ODIE corpus mentions had previously been annotated as `People`, `Procedure`, `DiseaseOrSyndrome`, `SignOrSymptom`, `Reagent`, `LaboratoryOrTestResult`, `OrganOrTissueFunction`, and `AnatomicalSite`; i2b2/VA corpus mentions as `Person`, `Problem`, `Treatment` and `Test`. In order to generalise the method across both corpora, and clinical narratives in general, these classifications were mapped

to two core types: `Person`, and the generic superclass '`Thing`' (i.e. in this case, clinical terms) (see Sections 8.3.8 and 8.3.9). The coreference component combines GATE ANNIE[26] text segmentation processing resources with custom named-entity annotators and integration plugins developed by the author to embed clinical domain knowledge and contextual cues into the text, in order to add semantic features to pronouns, `Person` and '`Thing`' mentions so that coreference relations can be computed.

The approach comprises five stages as shown in Fig. 8.3 and described in detail in Sections 8.3.3–8.3.9 below. In the examples presented, the text delimited by an annotation is shown in square brackets, the annotation type is shown in subscript in initial caps, and annotation features in subscript, lower case. JAPE patterns are shown in an abbreviated form, where token sequences are shown in square brackets, text in curly braces denotes the annotation name, and feature assignment statements are written as `Annotation.feature = value`.

8.3.3. Text segmentation

Standard GATE ANNIE[26] components provide initial shallow parsing and phrase chunking. Pattern-matching rules were written that split the source documents into sections and classify each, based on the text of identifiable headings (such as 'PHYSICAL EXAMINATION:' or 'LABORATORY DATA:') or paragraph content. Sections, sentences or paragraphs identified as being related to family history or historical lab data were then marked by the system as being potentially excluded from coreference of '`Thing`' mentions experienced by the patient in the current treatment episode (see Section 8.3.9).

8.3.4. Overview of coreference resolution approach

Coreference resolution rules follow similar heuristics to the multi-pass sieve recently presented by Lee et al.[6] for newswire text, but with specific consideration of world and clinical domain knowledge. While Lee et al. resolve pronouns on a final pass, we resolve pronominal coreference for each mention class first, and each potential mention-pair is considered only once, as described below. Furthermore, we address some of the weaknesses

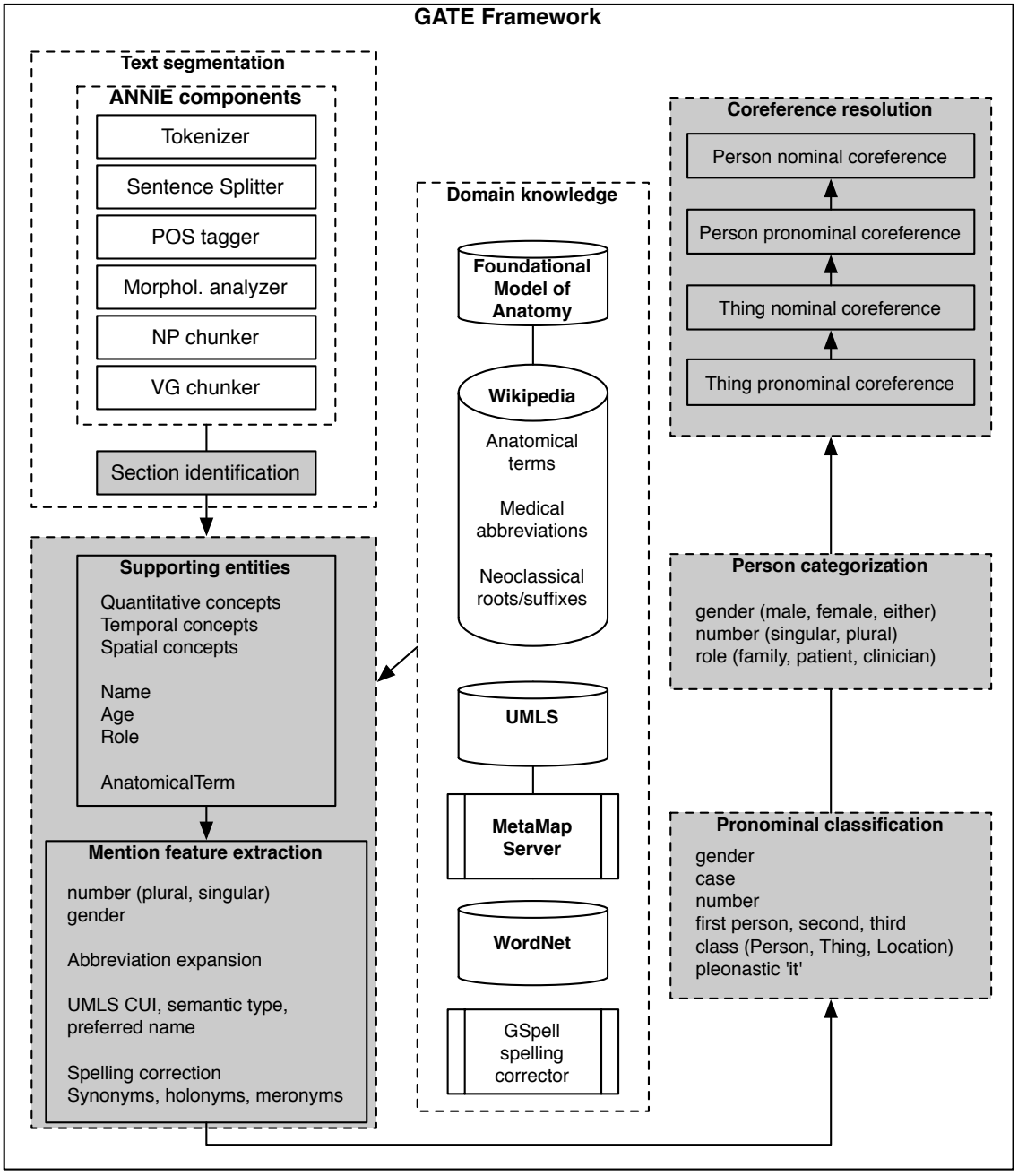


Figure 8.3.: Coreference architecture
'`Thing`' refers to a non-`Person` mention such as `Treatment`, `Test`, or `Problem`. Shaded areas represent components developed for this task; unshaded areas represent generic GATE components and external knowledge resources.
Source: Gooch & Roudsari[1]

of the traditional mention-pair approach, by making use of the contextual information surrounding each mention and/or pronoun, and by making use of *centering theory* to give preference to coreferents that grammatically agree with forward-looking centres[5].

Centering theory suggests that, in a coherent discourse, entities and their coreferent pronouns will occupy the same grammatical position in the sentence or clause – usually that of the subject where there is a single entity, but also in parallel subject/object pairs in the case of two or more entities and pronouns, as in:

[Patient]$_{\text{subject}}$ suffers from [lower back pain]$_{\text{object}}$.

[He]$_{\text{subject}}$ takes [Vicoprofen]$_{\text{indirect\_object}}$ for [this]$_{\text{object}}$.

Given two or more potential mentions that could be the correct antecedent based on gender and number, if the pronoun is the subject of its clause, then we select the mention that is also the subject of its own clause. This requirement is relaxed in the case of a single `Person` mention and a single personal pronoun in a clause, or two `Person` mentions and two personal pronouns, each of different genders.

We combine centering with the identification of the actors, or *protagonists*, who inform the narrative. Protagonist theory[27] suggests that narrative events are centred on one or more key actors. Coreferring actors share congruent verbs, and distinct sets of verbs are typically associated with different actor types. Narrative events can therefore be identified by a common protagonist and associated verbs[27]. By extension, we suggest that a known set of narrative events (e.g. the admission, assessment, test and treatment process documented in clinical notes) and associated verbs can be used to identify coreferring protagonists. This process is described in Sections 8.3.7 and 8.3.8.

Document traversal for generation of coreference chains

Taking the set of all mentions, we create subsets according to mention class, and within each subset, compare pairs of mentions in document order. For example, the first `Treatment` mention will need to be tested against all following `Treatment` mentions, the second against the third, fourth, etc. For a given subset, the maximum number of comparisons that need to be made, for each mention class, is given by

$$\sum_{i=1}^{n-1}(i-1) = \frac{n(n-1)}{2} \approx O(n^2)$$

where n is the number of mentions in the class.

However, for efficiency, each input subset is pruned of successful mention pairings during traversal, which will reduce the computational overhead of comparing large numbers of mentions if some of the mentions of each class are coreferential. This is illustrated in Figure 8.4, which represents a document containing two classes of mention, the selection of one class of mention and the coreference iteration process. As shown in the figure, when the candidate antecedent mention pointed to by the outer iterator matches a coreferent mention pointed to by the inner iterator, the features of the former are cloned to the latter, the outer iterator points to the coreferent mention, and the inner iterator is incremented to the next mention. Once the inner iterator completes, all coreferent mentions are pruned and the process repeats until the outer iterator completes. This process reduces the number of mention-pair tests from 21 ($n = 7 \Rightarrow n(n-1)/2 = 21$) to 11.

A set of linked lists corresponding to each coreference chain is thus created, where each mention is assigned a unique identifier and, for each link in the chain, we store the annotation id of the coreferent on the antecedent (and a back reference from the coreferent to the antecedent is created, to form a double-linked list). In doing so, the direction of the coreference relationship is preserved and the links in the chain of narrative events are made explicit.

Marking these links has the benefit of facilitating *in situ* evaluation via the GATE corpus quality assurance toolkit[25], which allows one to specify the feature names whose values must match between mentions of the same type and at the same position in the key set and system output. In this case, the coreference identifier is specified as the matching feature: the value (or null, for singletons) of the coreference identifier on each mention should therefore agree between the key set and system output. This coreference identifier functions as a pointer from the antecedent to the anaphor and thus considers the direction of the coreference relationship between terms. Precision and recall scores are then calculated as described in Chapter 4. This pairwise, directional measure attempts to avoid the anomalous results of the transitive metrics described above, although it is used here solely to give a snapshot evaluation of system performance. We still use the

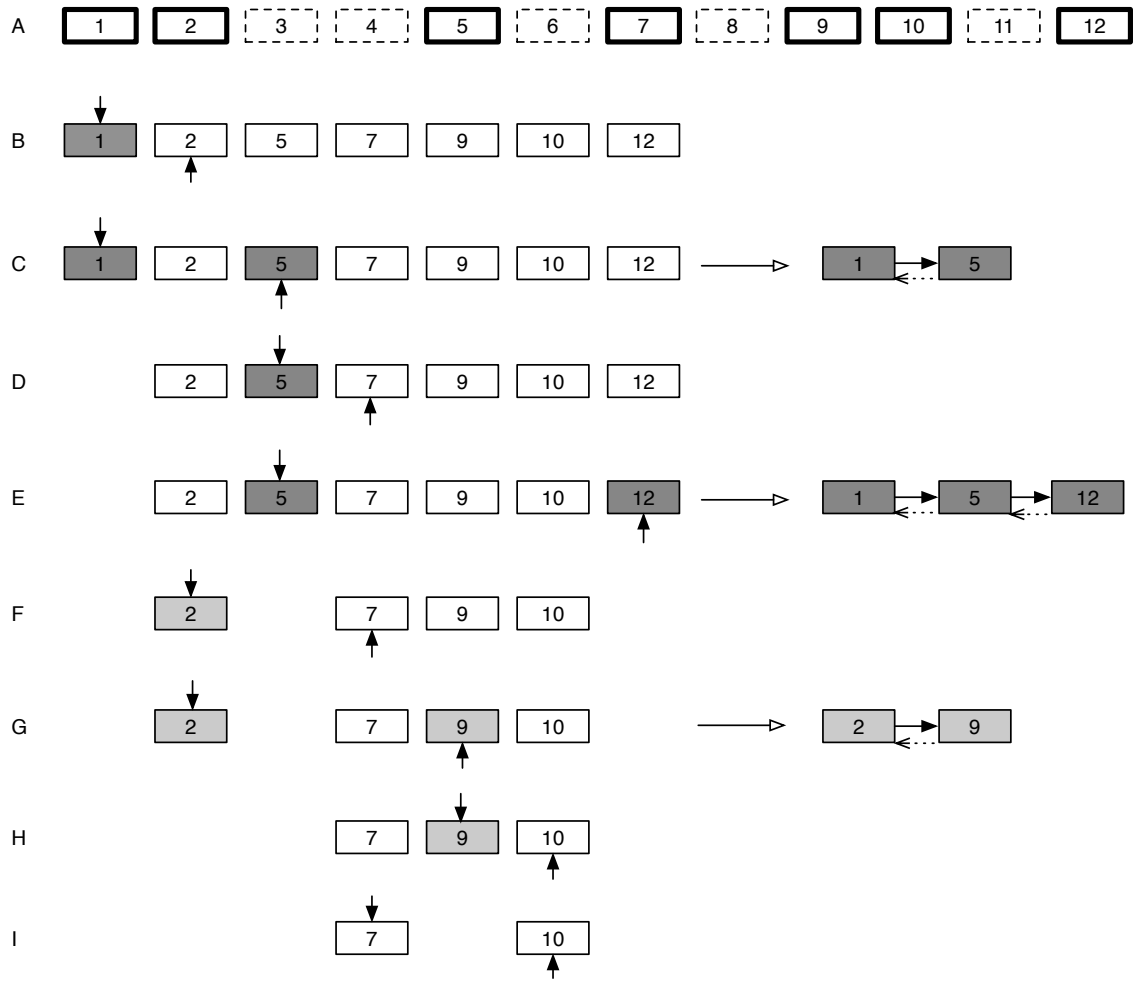


Figure 8.4.: Filtering and traversal of mention pairs with pruning
Boxes represent mentions; vertical arrows represent iteration pointers; short horizontal arrows represent coreference pointers; shading represents mention features. (A) Document containing 2 classes of mention, differentiated by bold and dashed outlines. (B) Filtering of mentions of the same class (mentions 1, 2, 5, 7, 9, 10, 12) and start of iteration. (C) Identification of coreferent mentions 1 and 5. When the mention pointed to by the outer iterator matches a mention pointed to by the inner iterator, the features of antecedent (1) are cloned to the coreferent (5) (shown as shading in the figure), and the corefence pair is created. The antecedent is then pruned (D), and the outer pointer then moves to the coreferent (5) and the inner iterator increments (7) for the next iteration. (E) Identification of coreferent mentions 5 and 12 and addition of 12 to the coreference chain. (F) The inner iterator has completed, which closes the coreference chain, the previous coreference pair are pruned and the iterators reset. (G) Identification of coreferent mentions 2 and 9 and creation of a new coreference chain. (H) Antecedent pruning and outer iterator moves to coreferent. (I) Inner iterator completes, closing previous coreference chain, pruning of previous coreferent and iterators reset.
Source: Gooch & Roudsari[1]

established metrics for evaluation of complete chains of coreference, as an average over all metrics should help balance out the strengths and weaknesses of individual measures, and

moreover this allows architecture-independent comparison with other systems, as discussed in [13].

Generating the unordered sets of coreference chains required by these metrics is, however, straightforward: each linked list is traversed in document order, and mentions and their start and end offsets are serialized to a coreference chain text file as described in Section 8.2. Before these coreference chains can be created, however, potentially co-referring mentions need to be identified. This process is described in detail in the following sections.

8.3.5. Identification of supporting entities, context and features

Determining whether two mentions corefer or not is usually dependent on the context in which those mentions appear, for example:

> [blood pressure]$_{\text{Test}}$ of [100/70]$_{\text{Measurement}}$ ... [blood pressure]$_{\text{Test}}$ of [140/80]$_{\text{Measurement}}$
>
> [Patient]$_{\text{Person}}$ has a past medical history of [hypertension]$_{\text{Problem}}$. [Patient's mother]$_{\text{Person}}$ also has [hypertension]$_{\text{Problem}}$.

The two '*blood pressure*' `Test` mentions do not refer to the same external event as they relate to different measurement events, and the two '*hypertension*' `Problem` mentions refer to conditions experienced by two different people.

Identification of entities and mention features that can be used to support or eliminate coreference between two mentions is a key task. Following Zheng et al.[8], selection of features and contextual cues was based on those used by general-purpose coreference systems, and which could be adapted to clinical texts, such as UMLS semantic type agreement, abbreviation expansion, plus additional features identified from a sample of documents from the training corpora.

Table 8.3 shows the supporting entities and features used for each mention class. To clarify, a supporting entity is a separate annotation identified by the system as providing information relevant to the context in which the `Person` or '`Thing`' mention appears. A supporting feature is something that is either already intrinsic to the mention itself (such as the headword of the noun phrase, or whether the word or phrase is singular or plural),

Table 8.3.: Supporting entities and features to identify mention context

Mention class	Supporting entity	Supporting feature
Person	Honorific, FirstName, Surname, GenderIdentifier, Age	role (family, patient, clinician) gender number
Pronoun	VG (verb group), IN (preposition)	gender number case
Thing	Section, Person, Date, Time, Duration, Number, Measurement, Frequency, MedicationRoute, AnatomicalTerm, SpatialConcept, TemporalConcept	number headword laterality (left, right, bilateral) normalizedString (abbreviation expansion, spelling correction, determiner removal) UMLS Concept Unique Identifier (CUI), UMLS preferred name, concept name, semantic type WordNet synonyms, hypernyms, holonyms, meronyms

or is the result of storing the text of a nearby supporting entity as a feature on the mention. For example,

[Mrs Smith]$_{\text{Person}}$, a [79-year-old] inpatient of Ward 1

The text '79-year-old' would be identified as an independent {Age} entity, but that supports the classification of the separate `Person` mention (see Section 8.3.7), whereas the gender of 'Mrs Smith' is a supporting feature, being an intrinsic property of the 'Mrs' honorific, and would be stored as a feature on the 'Mrs Smith' `Person` mention.

The general entity recognisers developed and described in Chapter 5 were used to identify supporting entities for context: quantitative and temporal concepts such as such as number, measurement, date, time, frequency, duration, and age. Anatomical and spatial concepts such as body locations and regions were identified using the anatomical term annotator developed from semantic decomposition of the Foundational Model of Anatomy[28] as described in Chapter 6 and which had previously been validated against

the ODIE corpus (see Results of Chapter 6).

The biomedical abbreviation annotator described in Chapter 7 was used to classify and expand abbreviations encountered in the text so that the expanded term could be processed by MetaMap (see below). A JAPE transducer was used to match abbreviations within mentions of the same class (so, for example, 'PT' as the content of a `Person` mention is more likely to mean '*physiotherapist*' or perhaps '*patient*', rather than '*prothrombin time*').

MetaMap[29] and the GATE `mmserver` integration plugin [30] (see Chapter 5) were used to identify term headwords and to add UMLS CUI and UMLS preferred names for each UMLS semantic type identified by MetaMap as features on each '`Thing`' mention. To reduce the number of features added, we used MetaMap's `-term_processing option` (i.e. each mention is treated as an atomic term for direct lookup against pre-coordinated entries in the Metathesaurus), only considered SNOMED CT mappings, and took only the highest-scoring MetaMap mapping group for each mention.

To correct misspellings, the biomedical spelling correction component developed using the GSpell API [31] (see Chapter 6) was used to provide in situ spelling suggestions for potentially misspelt terms in the clinical notes. The component adds to input mentions a feature containing the suggested correct spelling. To avoid false positives, spelling correction was limited to words longer than 3 characters, within an edit distance of 1, and only performed on mentions with no MetaMap mapping, and then a MetaMap re-match was attempted on the spell-corrected string.

A normalised string feature, generated from abbreviation expansion, spelling correction and removal of leading determiners and pronouns, was stored as the canonical form for each '`Thing`' mention (see Figure 8.5). The normalised string plus the mention's contextual features were used for the basis of mention-pair comparison (see Section 8.3.9).

To identify general synonyms, hypernyms and holonyms, a component that generates WordNet [32] annotations for given input mentions was developed. This was used to pass mention headwords and supporting entities (Table 8.3) to WordNet, and the output stored as features on the input mention (see Figure 8.6).

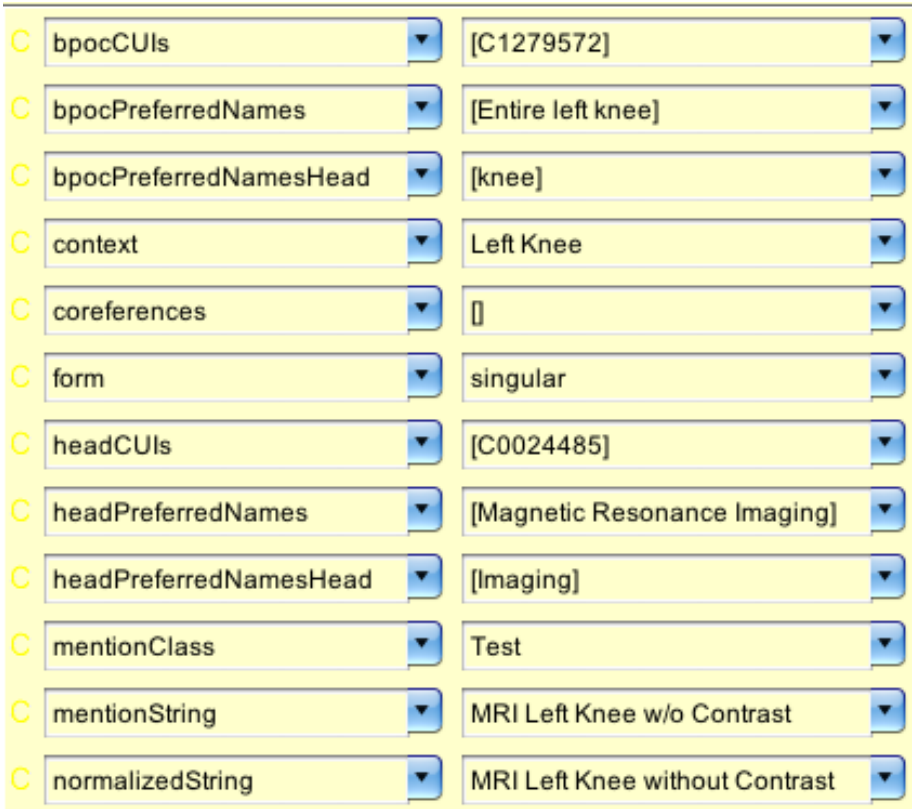


Figure 8.5.: String normalisation and contextual features

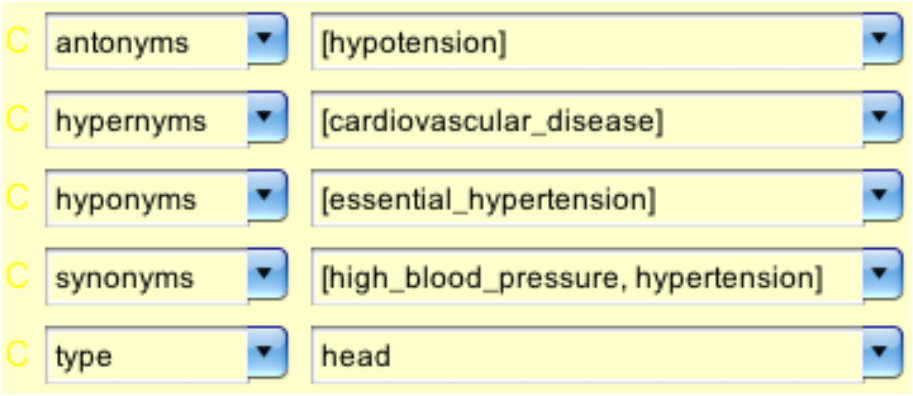


Figure 8.6.: Addition of WordNet synonyms and hypernyms

The surrounding context of each mention was identified by taking supporting entities within three Tokens either side of the mention, or within the mention itself, and storing this as a feature on the target mention. For example, given this input phrase:

$[\text{Culture}]_{\text{Test}}$ on blood sample was ... $[\text{Culture}]_{\text{Test}}$ on urine sample was ...

we obtain

[Culture]$_{\text{Test}}$ on [blood]$_{\text{AnatomicalTerm}}$ sample

⇒ `Mention`$_{\text{Test}}$`.anatomical_context = blood`

and

[Culture]$_{\text{Test}}$ on [urine]$_{\text{AnatomicalTerm}}$ sample

⇒ `Mention`$_{\text{Test}}$`.anatomical_context = urine`

For

MVA resulted in [3 broken left ribs]$_{\text{Problem}}$ and [1 broken right rib]$_{\text{Problem}}$

we obtain

[[3]$_{\text{Number}}$ broken [left]$_{\text{SpatialConcept}}$ [ribs]$_{\text{AnatomicalTerm}}$]$_{\text{Problem}}$

⇒ `Mention`$_{\text{Problem}}$`.spatial_context = left`,
`Mention`$_{\text{Problem}}$`.anatomical_context = rib`

and

[[1]$_{\text{Number}}$ broken [right]$_{\text{SpatialConcept}}$ [rib]$_{\text{AnatomicalTerm}}$]$_{\text{Problem}}$

⇒ `Mention`$_{\text{Problem}}$`.spatial_context = right`,
`Mention`$_{\text{Problem}}$`.anatomical_context = rib`

8.3.6. Pronoun classification

Using string matching and part-of-speech (POS) tags, we developed a general-purpose classifier in JAPE to categorise pronouns according to type (anaphoric or pleonastic); case: nominative (*I*, *he*, *she*); objective (*me*, *him*); possessive (*my*); reflexive (*myself*); nominative–possessive (*mine*, *hers*); number (singular or plural); class: `Person` (all personal pronouns), `Thing` (*it*, *that*, *these*, *those* – when not used in as determiners), `Location`

(*here*, *there*, *where*); person (first, second, third); and gender. Third-person plural pronouns (*they*, *their*, *them*) are provisionally categorised as `PersonOrThing` at this stage as their final assignment (`Person` or `Thing`) is context-dependent. The POS tagger was used to distinguish pronominal use of words such as 'these', 'those', 'that', 'this' from their use as determiners (e.g. '*wound closed with sutures ... [these]$_{PRP}$ will be removed*' vs. '*[these]$_{DET}$ sutures*') prior to the above classification.

Only anaphoric pronouns will participate in coreference, so pleonastic '*it*' and '*that*' references are identified using a set of general patterns that look for temporal phrases, verb 'to be' phrases ending in '*that*' or '*whether*' (e.g. '*It is unclear whether ...*', '*it is important to note that ...*') and modal 'to be' phrases ending in an infinitive or a preposition (e.g. '*It should be possible for ...*', '*It may be sensible to consider ...*'). JAPE expressions for these patterns, with accompanying examples in bold for clarity, are shown below (where `| ?` and `(n, m)` denote regular expression occurrence operators):

```
["It"]  {BE}  ({Day}|{Date}|{Time})
```

It is Tuesday
It was 10pm

```
["It"]  {VG.type == modal?, BE}
```

It is
It may be
It should be

```
({ADV}(0,2) {ADJ})(0,3)
```

somewhat unclear
important
possible

```
({VG.tense == Inf}?      ["whether|if|that"]) |  {IN}


                         whether
to note                  that
                                                for
```

where `VG` = verb group, `BE` = 'to be' verb group form, `ADV` = adverb, `ADJ` = adjective, `IN` = preposition, `Inf` = infinitive.

8.3.7. Person and personal pronoun categorization

Coreference systems for general English texts typically make use of gender, number and grammatical role information to resolve coreference of personal pronouns. A pseudo-pattern expressing possible pronominal coreference between a person and personal pronoun within the same sentence or between consecutive sentences might then be written as:

$\texttt{\{Mention\}}_{\text{Person,gender,number,grammar\_role}}$ $\texttt{(!\{Mention\}}_{\text{Person}}\texttt{)+}$

$\texttt{\{Mention\}}_{\text{Pronoun,gender,number,grammar\_role}}$

i.e. 'match a `Person` mention followed by a `Pronoun` mention where there are no intervening `Person` mentions', and where `Person.gender=Pronoun.gender`, `Person.number=Pronoun.number`, and `Person.grammar_role=Pronoun.grammar_role`.

For example:

> $[[\text{Jane}]_{\text{FirstName,female}}$ $[\text{Smith}]_{\text{Surname}}]_{\text{Person,female,singular,subject}}$ has a past history of [hypertension].
>
> $[\text{She}]_{\text{Person,female,singular,subject}}$ was admitted on ...

A typical system might also match occurrences of congruent name strings such as 'Smith', 'Jane', 'Ms Smith'. However, in anonymized clinical notes, the deidentification process potentially loses any link between the person's name and their gender, or between initial and subsequent mentions. Does 'XXXX' annotated as a `Person` refer to the patient, and are they male or female? Does Mr XYXY refer to the same person? Additional classification steps need to be employed to discriminate these cases.

For example, phrases extracted from the training corpus that identify the gender of the patient tend to be of the form:

[Patient]$_{\text{Person}}$ is a 40-year-old male with [type 2 diabetes]$_{\text{Problem}}$

[XXX]$_{\text{Person}}$ is an 80 y/o female admitted on ...

[This]$_{\text{Person}}$ is a baby boy born on ...

Which can be generalised to the pattern:

`{Mention} {BE}`

`{Age} {GenderIdentifier} ({VG} | {Mention}`$_{\text{Problem}}$`)`

⇒ `Mention.class=Person, Mention.semantic_role = patient, Mention.gender = GenderIdentifier`

From the preliminary analysis (see Section 8.3.1) of the 589 training documents, we established that the key protagonist in these clinical reports is the patient: 86% of all personal pronoun mentions referred to the patient, and 75% of all `Person` mentions also referred to the patient. The remaining mentions referred to members of the clinical team, to family/significant others, or to the person receiving the report. Therefore `Person` and personal pronoun mentions can be classified according to three main types:

1. patient
2. patient's family or significant other
3. clinician, which can be subcategorised as:
   - author
   - attending
   - receiver
   - referred clinicians (e.g. external teams, social workers etc)

Classification was performed using lexical rules and gazetteers of family relations (wife, daughter, brother, etc.), clinical roles and honorifics (physician, doctor, nurse, Dr., M.D.,

etc.) and contextual cues (e.g. section heading content and gender identifiers). Nominal `Person` mentions were classified as referring to the patient by default (as the patient is the key protagonist), unless the context suggested one of the other categories. For example, verb roots associated with a clinician include 'consult', 'attend', 'dictate', and certain past participles relate different protagonists, i.e.

`{Mention}`$_{\text{Person,semantic\_role1}}$ `["seen|treated|evaluated|treated ..."]`$_{\text{VG}}$ `["by"] {Mention}`$_{\text{Person,semantic\_role2}}$

`{Mention}`$_{\text{Person,semantic\_role1}}$ `{BE} ["referred|transferred ..."]`$_{\text{VG}}$ `["to"] {Mention}`$_{\text{Person,semantic\_role2}}$

⇒ `Mention.semantic_role1 = patient, Mention.semantic_role2 = clinician`

and

`{Mention}`$_{\text{Person,semantic\_role}}$ `["performed|signed|verified ..."]`$_{\text{VG}}$

⇒ `Mention.semantic_role = clinician`.

Or more generally, using role identifiers:

`{Mention}`$_{\text{Person,semantic\_role}}$ `{RoleIdentifier}`$_{\text{type}}$

⇒ `Mention.semantic_role = RoleIdentifier.type`

Information on the semantic role of the protagonist is also used to disambiguate `Person`-type abbreviations (see Chapter 7). For example:

The [pt]$_{\text{Person}}$ was referred to the [PT]$_{\text{Person}}$

Given the above protagonist role identifier patterns, both 'pt' and 'PT' are classified as `Person`, with the former given a 'patient' semantic role, the latter a 'clinician' role. This allows the abbreviations to be automatically expanded to 'patient' and 'physiotherapist', respectively, from the semantically typed lists of global abbreviations described in Chapter 7.

Personal pronouns were considered as having either global or local scope. By default, personal pronouns outside quoted speech have global scope. Second-person (*you*, *your*)

and third-person (*he*, *she*) singular pronouns are provisionally assigned to the patient if the pronoun's gender matches that of the patient. In the absence of gender cues, the document frequency of male and female pronouns were used to infer the patient's gender, given the prior probability (86%) that a personal pronoun refers to the patient. First-person pronouns (*I*, *we*, etc.) are assigned to the report's author.

Local scope exceptions are then identified as follows:

- A context switch triggered by a possessive pronoun, e.g. '*his wife ... she*', '*his oncologist ... he*'. Additionally, the locally scoped pronoun should agree in gender with that of the new context, if present.
- A context switch triggered by the appearance of a new actor, e.g. '*the social worker is Barbara Cole. She can be contacted on ...*' Again, gender features should agree, if present.
- Role of the report's receiver: The default protagonist is the patient, so references to *you*, *your*, etc. are assumed to be directed to the patient, unless it is clear that the recipient is a clinician (e.g. '*your patient*'), in which case, the second-person pronoun is assigned a clinical role.

8.3.8. Person coreference chain generation

Following the addition of the above-described classification features to `Person` and pronoun mentions, pairs of these mentions are traversed in document order and compared according to the following rules:

1. Strings are normalized by removing leading determiners and pronouns.
2. 'Who' pronouns are paired with the immediately preceding `Person` mention.
3. Pairs of nominal `Person`—third person-pronominal mentions are coreferenced if their genders (if present), scope, role/type and number (singular or plural) agree. In the absence of intervening plural '`Thing`' mentions, third-person plural pronouns

provisionally categorised as `PersonOrThing` were coreferenced with plural `Person` mentions with grammatical role agreement.

The features of the antecedent are cloned to the coreferent pronoun, so that the pronoun is now effectively a nominal `Person` mention, and the matching process continues from nominal to pronominal.

4. `Person` mentions classified as 'patient' are coreferenced if the genders agree. `Person` mention pairs classified as 'family' are coreferenced if the genders agree and the string values or WordNet synonyms agree (e.g. *sister* will corefer with *sibling*). Other `Person` mention pairs are coreferenced by evaluating the following, in order:

   a) Exactly matching name strings are coreferenced.

   b) Mentions with matching first names and surnames, where identifiable, are coreferenced.

   c) First-person pronouns of global scope are coreferenced and linked to the primary clinician (usually the report's author).

   d) Approximately matching strings over 4 characters long are coreferenced. Using the SecondString Java library[33], and following Cohen et al.[34] we take the mean value of the Jaro-Winkler[35] and Monge-Elkan [36] string comparison metrics, which returns a value between 0 (no match) and 1 (strong match). If the result exceeds a tuneable threshold, the two strings are coreferenced. This step allows de-identified name pairs such as ('***NAME[AAA, BBB]*', '***NAME[AAA]*'), and ('*Mr. BBBBB*', '*BBBB*') to be coreferenced. The threshold value was set at 0.85 – this was determined empirically by examining the Jaro-Winkler and Monge-Elkan scores on a small number (65) of randomly selected coreferent and non-coreferent mention pairs from the training set; values below 0.85 tended to accept false positives, values above 0.85 tended to reject true positives.

The following example demonstrates this process (square brackets denote the text spans of entities previously annotated as described in Section 8.3.5):

> [Mr WWWWW] is a [58 y/o] [gentleman] [who] was admitted ... by [Dr FFFF]. ... [He] was assessed by [Dr GGGGG] ... [She] has referred [WWW] to [the orthopedics team]; [he] will be followed up by [them].

Following the feature identification and classification described in Sections 8.3.7 and 8.3.6 above, we have

> [Mr WWWWW]$_{\text{Person,patient,male,singular}}$ is a [58 y/o]$_{\text{Age}}$ [gentleman]$_{\text{GenderIdentifier}}$ [who]$_{\text{Person}}$ was admitted ... by [Dr FFFF]$_{\text{Person,clinician,singular}}$. ...
>
> [He]$_{\text{Person,patient,male,singular}}$ was assessed by [Dr GGGGG]$_{\text{Person,clinician,singular}}$ ... [She]$_{\text{Person,female,singular}}$ has referred [WWW]$_{\text{Person,patient,male,singular}}$ to [the orthopedics team]$_{\text{Person,clinican,plural}}$; [he]$_{\text{Person,patient,male,singular}}$ will be followed up by [them]$_{\text{Person,plural}}$.

After steps 1–4 above, we have

> [*Mr WWWWW*]$_{\text{Person,patient,male,singular}}$ is a [58 y/o]$_{\text{Age}}$ [gentleman]$_{\text{GenderIdentifier}}$ [*who*]$_{\text{Person,patient,male,singular}}$ was admitted ... by [Dr FFFF]$_{\text{Person,clinician,singular}}$. ... [*He*]$_{\text{Person,patient,male,singular}}$ was assessed by [Dr GGGGG]$_{\text{Person,clinician,singular}}$ ... [She]$_{\text{Person,clinician,female,singular}}$ has referred [*WWW*]$_{\text{Person,patient,male,singular}}$ to [**the orthopedics team**]$_{\text{Person,clinican,plural}}$; [*he*]$_{\text{Person,patient,male,singular}}$ will be followed up by [**them**]$_{\text{Person,clinician,plural}}$.

where `Person` coreference chains are indicated via corresponding levels of emphasis: i.e. italics for the patient chain (*MrWWWWW* → *who* → *He* → *WWW* → *he*); underline for the clinician chain (Dr GGGGG → She); bold for the clinical team (**the orthopedics team** → **them**); singleton (Dr FFFF).

8.3.9. 'Thing' coreference chain generation

Coreference of general clinical terms follows a similar approach as for `Person` mentions. Anaphoric pronouns of class '`Thing`' (see Section 8.3.6) are resolved against the most recent '`Thing`' antecedent with the same grammatical role (e.g. subject, object, indirect

object), followed by the cloning of antecedent features to the anaphor so that the anaphor is converted to a nominal mention. Third-person plural pronouns provisionally classified as `PersonOrThing` (see Section 8.3.6 were coreferenced with plural '`Thing`' mentions '(e.g. '*the sutures ... they will be removed*') with grammatical role agreement in the absence of intervening plural `Person` mentions.

Nominal coreference is then attempted for pairs of mentions of the same class, in document order. This is more complex than for `Person` mentions and involves a voting process based on the number of matching features identified from rules given in the i2b2/VA coreference annotation guidelines[15], and the ODIE anaphoricity annotation guidelines[14]. In summary, these rules are:

For '`Thing`' mentions of the same class, consider pairing if:

1. mention synonyms (identified via UMLS or WordNet) refer to the same episode. For example, '*chills*' with '*shivering*' and '*inflammation*' with '*swelling*', if other contexts are equal;

2. a mentions occurs with its hypernym or a metonym (an alias identified via the term's preferred name in the UMLS or from WordNet) and if both refer to the same episode. For example, '*staph bacteraemia*' with '*the [infection]*$_{hypernym}$', '*stereotactic biopsy*' with '*the [procedure]*$_{hypernym}$', '*dyspnea*' with '*[shortness of breath]*$_{metonym}$', '*CABG*' with '*the [revascularization]*$_{metonym}$';

3. there is a holonym/meronym relation (identified via WordNet) between anatomical terms within or surrounding mentions;

4. there is agreement between the headwords of mention noun phrases where the antecedent is more specific than the coreferent, where all other contexts are equal. For example, '*intermittent right neck [swelling]*$_{headword}$' with '*the [swelling]*$_{headword}$'.

   Consider eliminating pairing where:

5. spatial concepts within each mention are different. For example, '*chronic [bilateral]*$_{SpatialConcept}$ *lower extremity swelling*' should not be coreferenced with '*the [right]*$_{SpatialConcept}$ *lower extremity swelling*';

6. the quantitative, temporal or anatomical context around each mention are different. For example:

   '*[2017-06-14 02:06AM]$_{TemporalConcept}$: WBC – 9.4*' vs.
   '*[2017-06-13 08:05PM]$_{TemporalConcept}$: WBC – 9.4*'

   '*blood pressure of [120/80]$_{Measurement}$*' vs. '*blood pressure' of [100/70]$_{Measurement}$*'.

   '*simple atheroma in the [aortic root]$_{AnatomicalTerm}$*' vs. '*simple atheroma in the [ascending aorta]$_{AnatomicalTerm}$*'.

7. Either mention is within a sentence or section of the document related to family history. Although family history information is important for decision support, as noted in Section 8.3.5, identical `Problem` strings with different experiencers are not in a coreference relationship.

   Coreferencing is not attempted if either of the mention pair occurs in an excluded section (rule 7) or if the contexts do not match (rules 5 and 6). A context match between mentions is made if there is a direct match between contextual features on both mentions (see Section 8.3.5) or there is a whole–part relation between the anatomical contexts of both mentions.

   If contexts match, or the antecedent mention has a contextual feature and the potential coreferent does not, then

   a) If there is an exact match between normalised strings (see Section 8.3.5), the coreference is marked and iteration continues with the next mention pair.

   b) Otherwise, consider marking a match if one or more of the following are true, in order of preference:

      i. The UMLS CUIs of the head word/phrase in each mention match, or if there is intersection between sets of head- word CUIs (where there is more than one), and the spatial contexts (e.g. left, right).

      ii. There is intersection between sets of anatomical terms within each mention and between sets of UMLS semantic types for the headword/phrase.

iii. The headwords and anatomical contexts match.

iv. There is an approximate string match, as measured by the mean Jaro-Winkler/Monge-Elkan score within the defined threshold (see Section 8.3.8).

8.3.10. Evaluation methodology

The above-described methods were implemented over five development iterations. For each iteration, experiments were performed against a weighted (according to the relative size of each corpus) random selection of 10 records from both training corpora. Performance was evaluated by analysing the results *in situ* using the simplified, pairwise coreference accuracy measure provided by the corpus QA tool within GATE and described in Section 8.3.4. The QA tool allows discrepancies between annotations and features in the training key set and the development set to be identified for each per-document in the corpus. These errors were inspected and adjustments to the feature extraction process were made where errors and their corresponding corrections could be generalised: for example, by making rules more (or less) specific and adjusting the scope of rules that were commonly misfiring. As generalisation of the method was one of the key aims, no document-specific changes to rules were made, although new abbreviations discovered in the training documents were used to enhance the abbreviation expansion component described in Chapter 7.

System validation against both the complete training set of 589 documents was then carried out both with the simplified, pairwise measure to give a snapshot of overall performance, and with the full evaluation metrics described in Section 8.3.4, in order to determine the accuracy of the complete, system-generated coreference chains.

When the 388 test documents became available from the i2b2/VA consortium, a final evaluation run was performed; this evaluation was also carried out independently in a recently published study, so that the results could be compared against other systems using the same metrics[13] Also using this evaluation data, a baseline evaluation system run of blind coreference was carried out where all mentions of the same class were linked into a single chain, as per Luo[23]. This run was compared against the contextually generated output using the Wilcoxon signed-rank test to determine whether the approach

offers significant improvements over the baseline.

Finally, the system was also evaluated both with and without the use of external domain knowledge resources (MetaMap, WordNet and GSpell) to assess the degree to which the use of these resources influenced the results, again compared using the Wilcoxon signed-rank test over matched document pairs. For '`Thing`' coreference, in the absence of these external resources, only rules that considered UMLS CUI, preferred name, synonym, meronym and hypernym matching should fail to fire as their input comparison sets would be empty of these domain knowledge features; all other aspects of the system – pronominal coreference, abbreviation expansion, spatial, temporal, quantitative and anatomical contexts, headword and approximate string matching – should remain as before. Exclusion of these external resources should also have a (small) impact on `Person` coreference as person-type synonyms (such as sister → sibling) will no longer be identified.

In all cases, *micro-averaged* precision and recall were calculated (see Chapter 4). In other words, the 'All classes' results are not simply the sum of each of the precision, recall and F_1-measure scores for each class over all documents divided by the number of classes (which is the macro-average), but are created by summing the true positives, false positives, false negatives and true negatives over all classes for each document, and then calculating the precision, recall and F_1-measures from these sums. Micro-average therefore gives the mean precision, recall and F_1-measure *per document*, which is useful in this case as it provides a measure of confidence in the system performance as a whole when processing a typical discharge summary or progress report (whereas macro-average would give a measure of performance for a typical mention class, which is perhaps less useful).

8.4. Results – training data

Summary validation results for the training portion of the i2b2/VA (492 documents) and ODIE corpora (97 documents) are reported in Tables 8.4 and 8.5 respectively. For each mention class, micro-average recall, precision and F_1-measure scores across the B^3, MUC and CEAF metrics output by the i2b2 coreference evaluation script, and the pairwise coreference identifier matching metric of the GATE corpus QA toolkit, are shown.

Table 8.4.: i2b2/VA training corpus coreference evaluation results (492 documents)

	Micro-average over i2b2/VA metrics*			Micro-average over GATE QA metrics**		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1
All classes	**0.905**	**0.855**	**0.878**	**0.923**	**0.923**	**0.923**
Person	0.886	0.880	0.883	0.917	0.920	0.917
Test	0.848	0.742	0.781	0.920	0.960	0.940
Treatment	0.867	0.775	0.813	0.897	0.927	0.913
Problem	0.862	0.788	0.820	0.870	0.900	0.887

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

** Results for 'All classes' account for singleton pronouns and thus differ from the mean over all classes shown

Table 8.5.: ODIE training corpus coreference evaluation results (97 documents)

	Micro-average over i2b2/VA metrics*			Micro-average over GATE QA metrics**		
	Precision	*Recall*	F_1	*Precision*	*Recall*	F_1
All classes	**0.771**	**0.828**	**0.796**	**0.765**	**0.765**	**0.765**
People	0.792	0.802	0.795	0.855	0.855	0.855
Disease	0.687	0.773	0.723	0.690	0.720	0.710
Symptom	0.802	0.782	0.791	0.730	0.760	0.745
Anat. Site	0.666	0.747	0.699	0.570	0.575	0.575
Reagent†	*0.352*	*0.160*	*0.131*	0.00	0.00	0.00
Organ Fn.	0.553	0.620	0.545	0.500	0.550	0.520
Lab. Result	0.798	0.711	0.740	0.800	0.875	0.835
Procedure	0.699	0.785	0.733	0.820	0.890	0.855

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

† null system results: treat italicised scores with caution.

** Results for 'All classes' account for singleton pronouns and thus differ from the mean over all classes shown

As shown in Table 8.5, in the case of the `Reagent` class, where the system did not identify any coreference relationships, the established metrics gave anomalous scores, consistent with the discussions in Section 8.2, whereas the null output was scored 0 by the pairwise metric in GATE as expected.

8.5. Results – test data

Tables 8.6 to 8.11 give a detailed breakdown of results by data source, class, and individual metric for baseline, knowledge-rich and knowledge-light system performance. Table 8.6 shows the micro-averaged precision, recall and F_1-measures for system coreference performance, with the inclusion of external domain knowledge, for the 322 documents from the three data centres (Beth Israel, Partners, UPMC) in the i2b2/VA test corpus. Scores for the B^3, MUC and CEAF metrics are shown, along with the average score over the three metrics.

Table 8.7 shows system results for the i2b2/VA test corpus where external domain knowledge is not used by the system assist in the identification of coreference relations. That is, Table 8.7 shows the effects of not adding UMLS or WordNet features to each mention: rules that match these features between prospective mention pairs will not have fired.

Table 8.8 shows baseline system results for the i2b2/VA test corpus. This table shows the effect of blind coreference between mention pairs of the same class linked into a single chain. That is, all `Person` mentions are joined into a chain, all `Test` mentions are joined into a separate chain, and so on.

Similarly, Tables 8.9 to 8.11 show the same three evaluations: system performance with domain knowledge (Tables 8.9), system performance without domain knowledge (Tables 8.10), and baseline system performance (Tables 8.11) for the 66 documents from the two data centres (Mayo Clinic, UPMC) in the ODIE test corpus. As noted in the footnotes to these tables, the B^3 and CEAF metrics gave anomalous scores for the `Lab Result` class when scoring system generated coreference chains against null output in the ground truth chains.

As can be seen in Tables 8.6 to 8.11, performance scores for a given class often varied greatly between evaluation metrics. This variation is consistent with findings by Cai and Strube[22]. As noted in Section 8.1, there is disagreement over which is the 'best' metric for evaluating coreference performance, so the mean scores across metrics have been used as a basis for the comparisons between system runs shown in Tables 8.12 and 8.13.

Table 8.6.: i2b2/VA test corpus coreference evaluation results, with external knowledge (322 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
With external domain knowledge resources (WordNet, MetaMap, GSpell)												
Beth Israel												
All classes	**0.941**	**0.968**	**0.954**	**0.849**	**0.751**	**0.797**	**0.886**	**0.828**	**0.856**	**0.892**	**0.849**	**0.869**
Person	0.964	0.933	0.948	0.932	0.941	0.937	0.770	0.826	0.797	0.889	0.900	0.894
Test	0.954	0.978	0.966	0.623	0.368	0.463	0.948	0.894	0.920	0.842	0.747	0.783
Treatment	0.909	0.952	0.930	0.778	0.623	0.692	0.858	0.788	0.821	0.848	0.788	0.814
Problem	0.907	0.949	0.928	0.766	0.619	0.685	0.866	0.789	0.826	0.846	0.786	0.813
Partners Healthcare												
All classes	**0.946**	**0.970**	**0.958**	**0.882**	**0.805**	**0.841**	**0.894**	**0.837**	**0.864**	**0.907**	**0.871**	**0.888**
Person	0.955	0.896	0.925	0.942	0.960	0.950	0.744	0.815	0.778	0.880	0.890	0.884
Test	0.941	0.97	0.955	0.706	0.444	0.545	0.942	0.888	0.914	0.863	0.767	0.805
Treatment	0.925	0.963	0.944	0.811	0.660	0.728	0.891	0.828	0.858	0.876	0.817	0.843
Problem	0.924	0.960	0.942	0.787	0.634	0.702	0.900	0.823	0.860	0.870	0.806	0.835
University of Pittsburgh Medical Center												
All classes	**0.942**	**0.960**	**0.951**	**0.858**	**0.790**	**0.822**	**0.873**	**0.817**	**0.844**	**0.891**	**0.856**	**0.872**
Person	0.913	0.880	0.896	0.919	0.912	0.915	0.718	0.695	0.706	0.850	0.829	0.839
Test	0.934	0.963	0.948	0.671	0.436	0.529	0.932	0.875	0.903	0.846	0.758	0.793
Treatment	0.929	0.937	0.933	0.675	0.626	0.650	0.871	0.842	0.857	0.825	0.802	0.813
Problem	0.921	0.958	0.939	0.774	0.581	0.664	0.899	0.804	0.849	0.865	0.781	0.817

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

Table 8.7.: i2b2/VA test corpus coreference evaluation results, no external knowledge (322 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
Without external domain knowledge resources												
Beth Israel												
All classes	**0.966**	**0.958**	**0.962**	**0.798**	**0.827**	**0.812**	**0.868**	**0.890**	**0.879**	**0.877**	**0.892**	**0.884**
Person	0.965	0.935	0.950	0.934	0.945	0.939	0.773	0.834	0.802	0.891	0.905	0.897
Test	0.968	0.976	0.972	0.538	0.485	0.510	0.945	0.942	0.943	0.817	0.801	0.808
Treatment	0.936	0.940	0.938	0.702	0.681	0.692	0.837	0.846	0.842	0.825	0.822	0.824
Problem	0.952	0.928	0.940	0.619	0.748	0.678	0.825	0.895	0.859	0.799	0.857	0.826
Partners Healthcare												
All classes	**0.970**	**0.957**	**0.963**	**0.826**	**0.871**	**0.848**	**0.867**	**0.897**	**0.882**	**0.888**	**0.908**	**0.898**
Person	0.955	0.896	0.925	0.942	0.960	0.951	0.745	0.818	0.780	0.881	0.891	0.885
Test	0.958	0.966	0.962	0.629	0.554	0.589	0.943	0.938	0.941	0.843	0.819	0.831
Treatment	0.948	0.950	0.949	0.718	0.727	0.722	0.861	0.881	0.871	0.842	0.853	0.847
Problem	0.962	0.938	0.950	0.593	0.760	0.666	0.846	0.915	0.879	0.800	0.871	0.832
University of Pittsburgh Medical Center												
All classes	**0.965**	**0.951**	**0.958**	**0.825**	**0.844**	**0.835**	**0.855**	**0.869**	**0.862**	**0.882**	**0.888**	**0.885**
Person	0.913	0.881	0.897	0.920	0.913	0.916	0.718	0.698	0.708	0.850	0.831	0.840
Test	0.951	0.960	0.955	0.609	0.530	0.567	0.932	0.922	0.927	0.831	0.804	0.816
Treatment	0.956	0.927	0.941	0.602	0.741	0.665	0.849	0.903	0.875	0.802	0.857	0.827
Problem	0.954	0.943	0.948	0.629	0.677	0.652	0.860	0.878	0.869	0.814	0.833	0.823

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

Table 8.8.: i2b2/VA test corpus coreference evaluation results, baseline (322 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
Baseline (blind coreference of mentions of the same class)												
Beth Israel												
All classes	**0.775**	**0.987**	**0.868**	**0.943**	**0.363**	**0.524**	**0.625**	**0.069**	**0.124**	**0.781**	**0.473**	**0.505**
Person	0.677	0.984	0.802	0.980	0.803	0.883	0.561	0.085	0.147	0.739	0.624	0.611
Test	0.170	0.986	0.290	0.772	0.068	0.125	0.248	0.006	0.011	0.397	0.353	0.142
Treatment	0.490	0.983	0.654	0.931	0.288	0.440	0.352	0.021	0.039	0.591	0.431	0.378
Problem	0.377	0.978	0.544	0.909	0.285	0.433	0.314	0.012	0.023	0.533	0.425	0.333
Partners Healthcare												
All classes	**0.732**	**0.987**	**0.841**	**0.950**	**0.414**	**0.577**	**0.669**	**0.095**	**0.166**	**0.784**	**0.499**	**0.528**
Person	0.542	0.990	0.700	0.992	0.855	0.919	0.798	0.131	0.225	0.777	0.659	0.615
Test	0.231	0.972	0.373	0.624	0.069	0.124	0.328	0.017	0.033	0.394	0.353	0.177
Treatment	0.263	0.987	0.415	0.944	0.282	0.434	0.403	0.022	0.042	0.537	0.430	0.297
Problem	0.150	0.979	0.260	0.895	0.252	0.393	0.310	0.012	0.022	0.452	0.414	0.225
University of Pittsburgh Medical Center												
All classes	**0.692**	**0.983**	**0.812**	**0.944**	**0.443**	**0.603**	**0.627**	**0.099**	**0.171**	**0.754**	**0.508**	**0.529**
Person	0.422	0.990	0.592	0.992	0.837	0.908	0.728	0.100	0.177	0.714	0.642	0.559
Test	0.213	0.969	0.349	0.647	0.083	0.147	0.339	0.022	0.041	0.400	0.358	0.179
Treatment	0.144	0.973	0.251	0.871	0.252	0.390	0.303	0.019	0.035	0.439	0.415	0.225
Problem	0.094	0.974	0.171	0.861	0.221	0.352	0.251	0.012	0.022	0.402	0.402	0.182

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

Table 8.9.: ODIE test corpus coreference evaluation results, with external knowledge (66 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
With domain knowledge (WordNet, MetaMap, GSpell)												
Mayo Clinic												
All classes	**0.931**	**0.868**	**0.898**	**0.737**	**0.852**	**0.790**	**0.616**	**0.797**	**0.695**	**0.761**	**0.839**	**0.794**
People	0.981	0.755	0.853	0.907	0.976	0.940	0.435	0.773	0.557	0.774	0.835	0.783
Disease	0.894	0.855	0.874	0.478	0.620	0.540	0.614	0.788	0.690	0.662	0.754	0.701
Symptom	0.921	0.947	0.934	0.745	0.623	0.679	0.823	0.791	0.807	0.830	0.787	0.807
Anat. Site	0.922	0.852	0.886	0.430	0.612	0.505	0.546	0.734	0.626	0.633	0.733	0.672
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.†‡	–	–	–	–	–	–	–	–	–	–	–	–
Lab Result†‡	–	–	–	–	–	–	–	–	–	–	–	–
Procedure	0.937	0.876	0.905	0.469	0.692	0.559	0.745	0.848	0.793	0.717	0.805	0.752
University of Pittsburgh Medical Center												
All classes	**0.906**	**0.879**	**0.892**	**0.796**	**0.840**	**0.817**	**0.598**	**0.723**	**0.654**	**0.767**	**0.814**	**0.788**
People	0.885	0.842	0.863	0.894	0.919	0.906	0.447	0.560	0.497	0.742	0.774	0.755
Disease	0.894	0.823	0.857	0.564	0.694	0.622	0.589	0.757	0.662	0.682	0.758	0.714
Symptom	0.914	0.950	0.932	0.742	0.687	0.713	0.874	0.867	0.870	0.843	0.835	0.838
Anat. Site	0.860	0.842	0.851	0.697	0.736	0.716	0.538	0.644	0.586	0.698	0.741	0.718
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.	0.750	0.833	0.789	0.000	0.000	0.000	0.528	0.792	0.633	0.426	0.542	0.474
Lab Result‡	*0.881*	*1.000*	*0.937*	*0.000*	*0.000*	*0.000*	*0.923*	*0.706*	*0.800*	*0.601*	*0.569*	*0.579*
Procedure	0.883	0.762	0.818	0.609	0.778	0.683	0.577	0.797	0.670	0.690	0.779	0.724

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

† 0 system results;

‡ 0 ground truth results; treat scores with caution.

Table 8.10.: ODIE test corpus coreference evaluation results, no external knowledge (66 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
Without external domain knowledge resources												
Mayo Clinic												
All classes	**0.954**	**0.852**	**0.900**	**0.692**	**0.881**	**0.775**	**0.582**	**0.810**	**0.677**	**0.743**	**0.848**	**0.784**
People	0.981	0.755	0.853	0.907	0.976	0.940	0.435	0.773	0.557	0.774	0.835	0.783
Disease	0.926	0.843	0.883	0.389	0.656	0.488	0.584	0.803	0.676	0.633	0.767	0.682
Symptom	0.949	0.917	0.933	0.569	0.690	0.624	0.753	0.841	0.795	0.757	0.816	0.784
Anat. Site	0.940	0.836	0.885	0.347	0.594	0.438	0.505	0.738	0.600	0.597	0.723	0.641
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.†‡	–	–	–	–	–	–	–	–	–	–	–	–
Lab Result†‡	–	–	–	–	–	–	–	–	–	–	–	–
Procedure	0.948	0.858	0.901	0.346	0.692	0.461	0.695	0.859	0.768	0.663	0.803	0.710
University of Pittsburgh Medical Center												
All classes	**0.930**	**0.854**	**0.890**	**0.748**	**0.864**	**0.801**	**0.546**	**0.751**	**0.633**	**0.741**	**0.823**	**0.775**
People	0.885	0.842	0.863	0.894	0.919	0.906	0.447	0.560	0.497	0.742	0.774	0.755
Disease	0.925	0.777	0.845	0.444	0.756	0.559	0.517	0.802	0.629	0.629	0.778	0.678
Symptom	0.935	0.931	0.933	0.565	0.729	0.636	0.781	0.892	0.833	0.760	0.851	0.801
Anat. Site	0.904	0.808	0.853	0.591	0.765	0.667	0.474	0.704	0.566	0.656	0.759	0.695
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.	0.750	0.833	0.789	0.000	0.000	0.000	0.528	0.792	0.633	0.426	0.542	0.474
Lab Result‡	*0.881*	*1.000*	*0.937*	*0.000*	*0.000*	*0.000*	*0.923*	*0.706*	*0.800*	*0.601*	*0.569*	*0.579*
Procedure	0.901	0.721	0.801	0.493	0.791	0.607	0.505	0.807	0.621	0.633	0.773	0.676

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

† 0 system results;

‡ 0 ground truth results; treat scores with caution.

Table 8.11.: ODIE test corpus coreference evaluation results, baseline (66 documents)

	Metric											
	B^3			**MUC**			**CEAF**			**Avg***		
	Prec	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*	*Prec*	*Rec*	*F*
Baseline (blind coreference of mentions of the same class)												
Mayo Clinic												
All categories	**0.737**	**0.941**	**0.827**	**0.911**	**0.646**	**0.756**	**0.581**	**0.273**	**0.372**	**0.743**	**0.620**	**0.652**
People	0.779	0.959	0.860	0.987	0.940	0.963	0.796	0.374	0.509	0.854	0.758	0.777
Disease	0.702	0.929	0.800	0.732	0.424	0.537	0.551	0.239	0.333	0.662	0.531	0.557
Symptom	0.604	0.974	0.746	0.843	0.326	0.470	0.591	0.156	0.247	0.679	0.485	0.488
Anat. Site	0.670	0.960	0.789	0.909	0.498	0.643	0.656	0.256	0.369	0.745	0.571	0.600
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.†‡	–	–	–	–	–	–	–	–	–	–	–	–
Lab Result‡	*0.625*	*1.000*	*0.769*	*0.000*	*0.000*	*0.000*	*0.833*	*0.833*	*0.833*	*0.486*	*0.611*	*0.534*
Procedure	0.634	0.917	0.750	0.679	0.371	0.480	0.644	0.251	0.361	0.652	0.513	0.530
University of Pittsburgh Medical Center												
All categories	**0.772**	**0.929**	**0.843**	**0.887**	**0.685**	**0.773**	**0.488**	**0.267**	**0.345**	**0.716**	**0.627**	**0.654**
People	0.823	0.891	0.856	0.929	0.845	0.885	0.398	0.262	0.316	0.717	0.666	0.686
Disease	0.669	0.892	0.765	0.752	0.513	0.610	0.457	0.193	0.271	0.626	0.533	0.549
Symptom	0.575	0.968	0.721	0.839	0.325	0.468	0.610	0.123	0.205	0.675	0.472	0.465
Anat. Site	0.678	0.931	0.785	0.894	0.654	0.755	0.567	0.254	0.351	0.713	0.613	0.630
Reagent†‡	–	–	–	–	–	–	–	–	–	–	–	–
Organ Fn.	0.736	0.889	0.805	0.500	0.333	0.400	0.583	0.583	0.583	0.606	0.602	0.596
Lab Result‡	*0.464*	*1.000*	*0.634*	*0.000*	*0.000*	*0.000*	*0.594*	*0.210*	*0.310*	*0.353*	*0.403*	*0.315*
Procedure	0.730	0.873	0.795	0.797	0.640	0.710	0.711	0.384	0.499	0.746	0.632	0.668

* Unweighted average of MUC, B^3 and CEAF scores according to i2b2/VA evaluation script[16].

† 0 system results;

‡ 0 ground truth results; treat scores with caution.

Tables 8.12 and 8.13 show the results of Wilcoxon signed-rank tests across matched pairs between system and control, where control was 1) system without external domain knowledge resources and 2) baseline of blind coreference of mentions of the same class. Two-tailed tests were performed as no prior assumptions were made on which approach would be an improvement over the other; W is the sum of the signed ranks: a positive value indicates improved system results over the control, a negative value indicates lower system results than the control.

As shown in Table 8.12, on the i2b2/VA corpus the system performed significantly better than the baseline ($p < 0.01$) in terms of precision, recall and F_1-measure. The use of the external knowledge bases led to a significant increase in precision ($p < 0.01$) but with significantly reduced recall ($p < 0.01$) – the reduction in which outweighed the increase in precision and led to a significant decrease in overall performance as recorded in the reduction in F_1-measure.

Similarly, as shown in Table 8.13, for the ODIE corpus, in terms of F_1-measure the system performed significantly better than the baseline ($p < 0.05$), although precision in comparison to the baseline was not significantly improved ($p > 0.05$). As with the i2b2 corpus, the use of external domain knowledge significantly increased precision ($p < 0.01$) but with borderline-significantly reduced recall ($p = 0.05$). The effect of the slight recall reduction was not so great as to outweigh the effect of increased precision; overall, and in contrast to the i2b2 corpus, domain knowledge led to significantly increased F_1-measure performance ($p < 0.01$) for the ODIE corpus.

Table 8.12.: Two-tailed Wilcoxon signed-rank tests for system performance over baseline and with or without domain knowledge, i2b2/VA corpus (322 documents, 15 classes)

Wilcoxon $W, n_1 = n_2 = 15$, two-tailed	Precision	Recall	F_1
With external domain knowledge vs. baseline	$W = 120, p < 0.01$	$W = 120, p < 0.01$	$W = 120, p < 0.01$
With external domain knowledge vs. without	$W = 99, p < 0.01$	$W = -120, p < 0.01$	$W = -113, p < 0.01$

Table 8.13.: Two-tailed Wilcoxon signed-rank tests for system performance over baseline and with or without domain knowledge, ODIE corpus (66 documents, 18 classes)

Wilcoxon $W, n_1 = n_2 = 18$, two-tailed	Precision	Recall	F_1
With external domain knowledge vs. baseline	$W = 6, p = 0.865$	$W = 88, p < 0.01$	$W = 78, p < 0.05$
With external domain knowledge vs. without	$W = 55, p < 0.01$	$W = -39, p = 0.05$	$W = 55, p < 0.01$

8.5.1. Coreference chain lengths

In the ground truth test data, the number of mentions, chains, mean and maximum coreference chain lengths were 3002, 419, 5.7 and 90 for the ODIE corpus; and 43,867, 5277, 4.3 and 122 for the i2b2/VA corpus. The mean number of true mentions, chains and coreference relations per document were 45.5, 6.4 and 36.2 for the ODIE corpus, and 136.3, 16.4 and 70.5 for the i2b2/VA corpus. Selecting a mention at random, and assuming all mentions have an equal chance of participating in coreference, the likelihood that a given mention will appear in a coreference chain can be estimated as

$$p_c = N_{chains} * \mu_c / N_{mentions}$$

where N_{chains} is the total number of true chains in the ground truth data, μ_c is the mean chain length and $N_{mentions}$ the total number of true mentions. For the ODIE corpus, this gives $p_c \approx 0.8$, for the i2b2/VA corpus $p_c \approx 0.5$.

Coreference chain length varied widely between document type: discharge and progress reports from both corpora had higher mean (5.42) and maximum chain length (106) than radiology, surgery and pathology reports (mean 3.61, maximum 18).

8.5.2. Pronoun distribution

As shown in Figure 8.7, the distribution of pronouns in the test set was similar to that of the training set (see Section 8.3.1): third person singular personal pronouns dominated (black bars in Figure 8.7), and the majority of these also participated in coreference (white bars

in Figure 8.7), whereas the majority of 'this', 'that' and 'it' pronouns did not – probably because their use was largely pleonastic (as discussed in Section 8.1, or as determiners. As shown in Figure 8.7, while there was generally close agreement between the ground truth and system in terms of which pronouns participated in a coreference chain, the system over-generated 'this' coreferences and under-generated 'they' coreferences.

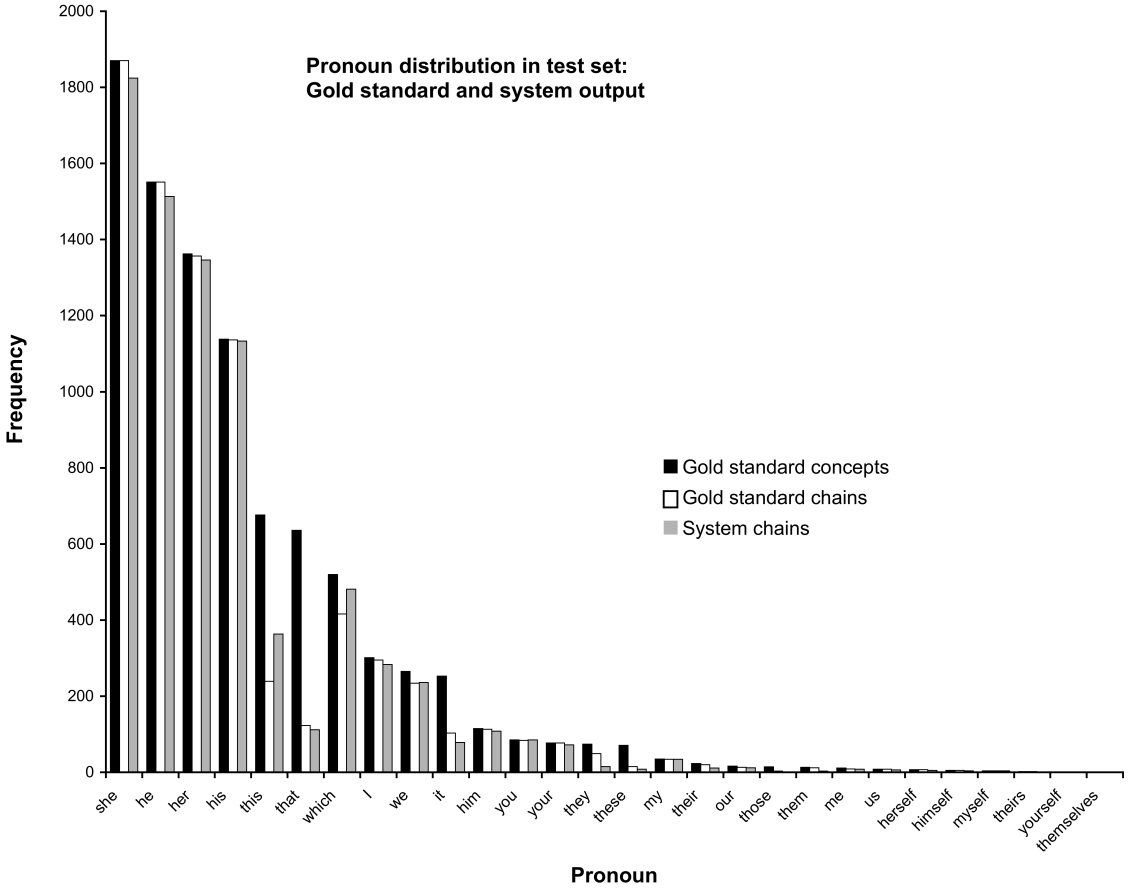


Figure 8.7.: Distribution of pronouns in the test set

8.6. Error analysis and discussion

Overall performance on both the training and test data was in close agreement (79.6% vs. 79.1% for the ODIE corpus; 87.8% vs. 87.6% for the i2b2/VA corpus), which suggests that the rules for feature extraction and coreference resolution were not over-fitted to

the training set. Comparison of the system output (with domain knowledge) against the ground truth suggested four categories of discrepancy:

1. *Errors of commission or omission*: For `Person` mentions, these resulted from incorrect categorisation by the system. For other classes, errors occurred where contextual cues had been incorrectly identified, or where the string similarity metrics had reported a false match or lack of match. Spurious pronominal coreferences occurred where pleonastic '*it*' and '*that*' pronouns had been incorrectly classified as anaphoric.

2. *Split coreference chains*: coreference relations were correct, but were reported across 2 or more chains, when a single chain should have been reported.

3. *Concatenated coreference chains*: the coreference chain was partially correct, but contained mentions incorrectly coreferenced with other chained concepts that should have appeared in a separate chain.

4. *Deterministic behaviour*: Unlike machine learning approaches, deterministic rules cannot model inconsistencies in the ground truth data. In 28 of the 46 Beth Israel records in which the attending physician was annotated, the 'Attending' heading and physician name following were coreferenced. In the remaining 18, they were not. There were other inconsistencies in the coreferencing of names with their clinical role in both corpora. However, our deterministic rules did not allow for such inconsistencies, and always coreferenced physician names with their clinical role.

In the following sections, we discuss these factors in more detail by examining the differences in system output when using external domain knowledge resources versus system output without domain knowledge, in comparison to the ground truth data.

8.6.1. Effects of domain knowledge

As noted in Tables 8.12 and 8.13, the use of domain knowledge led to a surprising drop in overall system performance when processing the i2b2/VA data (which consists solely of discharge summaries and progress notes), whereas with the ODIE data (which consists

largely of surgical, pathology and radiology reports) domain knowledge did improve system performance as expected. To investigate the effects of domain knowledge on coreference resolution in more detail, a random sample of 10 documents from each corpus was taken and the differences between the system coreference chain output (see Section 8.1) with domain knowledge vs. without were identified using the `diff` utility. These differences were compared against the ground truth data. The results of this analysis are summarised in Table 8.14.

Table 8.14.: Analysis of impact of domain knowledge resources

Report type	**Domain knowledge effects**	**Comparison with ground truth**
i2b2/VA corpus		
Progress	Adds '*pulmonary hypertension*' to '*pulmonary fibrosis*' chain	Absent
	New chain '*bilateral small pleural effusions*' → '*small bilateral effusion*'	Absent; different contexts: first relates to a lung finding, the second preceded by '*heart appears enlarged*'
Discharge	Adds '*back pain*' → '*chronic pain*' → '*pain*' to '*chronic back pain*' chain and concatenates '*breakthrough pain*' chain into this chain	Present but '*breakthrough pain*' is in a separate chain
Discharge	Adds '*hcv cirrhosis*' to '*hep C cirrhosis*' chain	Present
	Adds '*hepatic encephalopathy*' to '*encephalopathy*' chain	Present

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	Adds '*increased lethargy*' to '*lethargy*' chain	Present
	Adds '*worsening confusion*' to '*confusion*' chain	Present
	Adds '*any pain*' to '*any abdominal pain*' chain	Absent, but system may be correct: both relate to patient's later denial of pain following admission for abd. pain
	New chain '*5 cores of varices*' → '*massive peri splenic varices*'	Present, but also has '*esophageal varices*'
	New chain '*hypertension*' → '*severe portal hypertension*'	Absent; but system could be correct: first is from patient's history, second is ultrasound finding on admission
Discharge	New chain '*some fullness in the suprapatellar pouch*' → '*calcification densities in the suprapatellar bursa region*'	Absent; but system may be correct: first is presenting condition, second is MRI confirmation
	New chain '*right knee aspirations*' → '*the right knee examination*'	Absent; former is historical procedure, latter is procedure in current admission
Progress	New chain '*repair*' → '*his SMA repair*'	Present
Discharge	New chain '*epigastric pain*' → '*abd pain*'	Absent; but system may be correct as contexts are the same

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
Discharge	New chains '*an l4 decompressive lumbar laminectomy*' → '*l4 decompressive lumbar laminectomy*'; and '*surgical intervention*' → '*all surgery*'	Present
	New chain '*neuroimaging studies*' → '*the studies*'	Absent; but system is correct: clear definite descriptor
	New chain '*symptoms*' → '*any worsening symptoms*'	Absent; but system may be correct: the context implies that the second mention refers to the former, admission symptoms
	Adds '*worsening low back pain*' to '*pain*' chain	Present
	Adds '*any discoloration around the incision line*' to '*incision line*' chain	Absent; but context suggests system may be correct
Discharge	New chain '*hypertension*' → '*mild secondary pulmonary arterial hypertension*'	Absent, but system probably correct: first is discharge diagnosis in summary section, second is finding leading to the summary diagnosis
	New chain '*increasing dyspnea*' → '*any symptoms of shortness of breath*' → '*shortness of breath*'	Absent, but system may be correct: first is admitting symptom, latter are patient's later denial of these symptoms

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	New chain '*the transesophageal echo images*' → '*post-closure images*'	Absent; former is pre-closure, latter is post-closure
	New chain '*tamponade*' → '*obvious intracardiac shunting*'	Absent; but system is possibly correct, or at least useful, as tamponade may be a complication of shunting
	New chain '*the sheath*' → '*the sheaths*'	Present
	Adds '*a 22 - mm amplatzer atrial septal defect occluder*' → '*an 8 - mm atrial septal defect occluder*' → '*a 26 - mm amplatzer atrial septal device occluder*' → '*occluder device in the atrial septum*' to '*asd closure*' procedure chain	Present, but in separate chain (device and procedure should not be coreferenced)
	Adds '*significant left to right flow*' to '*prominent left to right shunting*' chain	Present
	Adds '*a left to right shunt at the atrial level*' to '*residual shunt*' chain	Absent, should be in '*prominent left to right shunting*' chain

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	Adds '*the 22 - mm device*' → '*the smaller device*' → '*the closure device*' to the '*both devices*' chain	Absent; should be in separate chain
	Adds ''*chest pressure*' → '*chest discomfort*' → '*sudden pain at that site*' to '*chest pain*' chain	Absent; prospective symptoms ('please report if you have ... ')
	Adds '*a repeat MRI*' to the '*an MRI*' chain	Present, but should have been attached to '*a cardiac MRI*' chain
Discharge	New chain '*chest x-ray*' → '*ct of chest*'	Absent
	Adds '*the patient*'s left back pain' → '*the back pain*' to '*left back pain*' chain	Present
	Adds '*multiple bilateral pulmonary nodules*' → '*disease recurrence*' to '*COPD*' chain	Absent; first mention is in separate chain
Discharge	Adds '*enlarged ovary*' → '*a full right ovary*' → '*ovarian cancer*' to '*ovarian mass*' chain	Absent; but system may be partially correct although last 2 mentions may belong in separate chain
	Adds '*several years urinary incontinence*' to '*incontinence*' chain	Present

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	Adds '*the procedure*' to '*birch procedure*' chain	Absent; not clear to which procedure the latter refers
	Adds '*jackson-pratt drain placement*' to '*suprapubic bladder catheter placement*' chain	Absent
	New chain '*abdominal pain*' → '*pain*'	Absent; former refers to historical event
	New chain '*the pelvic examination*' → '*the rectovaginal examination*'	Absent
ODIE corpus		
Progress	New chain '*possibly dysarthria*' → '*mild dysarthria*'	Present, but also has '*a thick tongue*' at start of chain
	New chain '*anxiety attack*' → '*TIA*'	Absent; TIA expanded to '*transient ischaemic attack*' leading to erroneous headword match
Progress	Adds '*the humerus*' → '*a larger humeral head*' to '*humeral head*' chain	Present; but as part of a larger chain that includes '*glenoid cartilage*' → '*the shoulder*' → '*rotator cuff*', whereas system has these in separate chains
	Adds '*the cartilage of the glenoid*' to '*the glenoid*' chain	Present, but should be in the above, longer chain
	Adds '*surgical treatment*' to '*surgery*' chain	Present

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	Adds '*two surgical procedures*' → '*half a dozen surgical procedures*' to '*surgery*' chain	Absent; these refer to previous procedures, not ones in current care episode
	New chain '*pain*' → '*left shoulder pain*' → '*significant pain in the shoulder*'	Present, but also has '*discomfort*' mention missed by system chain
Progress	Adds '*mild chest tightness*' → '*this chest pain*' to '*chest discomfort*' chain	Present, but also has '*the tightness*' and '*a tightness*' missed by system chain
	Adds '*disrupted snoring*' to '*snoring*' chain	Present
	Adds '*coronary artery disease*' to '*gastroesophageal reflux*' chain	Absent
	Adds new chain '*post-prandial angina*' → '*angina*'	Present
	Adds new chain '*allergic rhinitis*' → '*vasomotor rhinitis*'	Absent; not strictly coreferential although may be useful to link them
Radiology	Adds new chain '*xray knee*' → '*radiographs of the left knee*'	Present
	Adds '*bilaterial knees*' → '*both knees*' to '*the left knee*' chain	Present

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
	Adds '*right knee*' to '*the left knee*' chain	Absent; system cascading error resulting from addition of '*both knees*' to this chain
Pathology	Adds new chain '*the closest resection margin*' → '*the mucosal resection margins*'	Present
Surgical-pathology	Adds '*polyp*' → '*ascending colon polyp*' to '*colon polyp*' chain	Present
	Removes '*colon polyp ascending*' from '*colon polyp*' chain	Present, has '*colon polyp ascending*' missing from system chain
Surgical-pathology	Adds '*8 cc of cloudy yellow pleural fluid*' to '*pleural fluid*' chain	Present
	Adds '*hypocellular fluid*' → '*8cc cloudy yellow fluid*' to '*pleural fluid*' chain	Absent; in separate chain, but system may be correct here
Pathology	Adds new chain '*sigmoid colon*' → 'colon, sigmoid'	Present, but also has '*the nearest resection margin*' → '*muscularis*' → '*Surgical margins*' missing from system chain
	Adds new chain '*the overlying subserosal adipose tissue*' → '*subserosal*'	Absent; overlapping annotation error

Continued on next page

Table 8.14 – continued from previous page

Report type	Domain knowledge effects	Comparison with ground truth
Radiology	Adds 'an unenhanced ct of the chest, abdomen and pelvis' to '*ct abdomen*' chain	Present, but also has '*this non-contrast CT examination*' missing from system chain
	Adds new chain '*a mild to moderate size left-sided pneumothorax*' → '*mild to moderate sized left-sided pneumothorax*'	Present
	Adds new chain '*the collecting system of both kidneys*' → '*the bilateral collecting systems*'	Present
Pathology	Adds new chain '*ascending mass*' → '*right rectus sheath mass*' → 'a 2 x 1.5 x 1 cm mass'	Present, but '*ascending mass*' is in separate chain
	Adds new chain 'infiltrating grade 3 (of 4) adenocarcinoma' → 'invasive grade 3 (of 4) mucinous adenocarcinoma'	Absent, may refer to different masses
	Adds new chain '*splenic flexure polyp*' → '*hyperplastic polyp*'	Present
	Adds new chain '*soft tissue*' → '*fibrous tissue*'	Absent, latter refers to a hepatic module

As expected, and as shown in Table 8.14, domain knowledge had no effect on pronominal coreference, only on nominal coreference. That is, domain knowledge did not lead to the addition or removal of any mention–pronoun coreference pairs to the coreference chains in

any of the documents sampled.

In the 10 documents sampled from the i2b2/VA corpus, we can see from Table 8.14 that, in comparison to system output without domain knowledge, domain knowledge added 24 new chains, or additions to existing chains, that were absent from the ground truth. However, domain knowledge added only 15 new chains, or additions to existing chains, that were also present in ground truth. In contrast, in the 10 documents sampled from the ODIE corpus, domain knowledge added only 9 new chains, or additions to existing chains, that were absent from the ground truth, but added 20 new chains, or additions to existing chains, that were also present in the ground truth.

These results initially suggest that domain knowledge generally adds more incorrect than correct coreference relations to discharge summaries and progress reports, but adds more correct than incorrect relations to surgical, pathology and radiology reports. Why might this be the case? As noted in the third column of the table, in the 24 false-positives generated by inclusion of domain knowledge in the i2b2/VA discharge/progress reports, at least 10 were most likely valid coreference relations. Overall, in the discharge summaries/progress reports, domain knowledge had a variety of somewhat confounding effects: picking up valid relationships that may or may not be present in the ground truth, sometimes leading to previously valid chains being joined by invalid chains, or splitting longer chains; and adding invalid relationships due to context not being adequately identified. However, in the ODIE surgical-pathology/radiology reports, the addition of domain knowledge seemed to have a more consistent effect: a much greater proportion of the new chains and relationships added by the inclusion of domain knowledge matched the ground truth data, and most of the false positives (7 out of the 9) appeared, on inspection, to be valid false positives.

Discharge and progress notes documents were typically much longer than the lab reports, so perhaps it is not surprising that potentially valid relations were missed during manual annotation of these longer documents. Also the relationships identified in the discharge summaries by the manual annotators are generally quite straightforward, whereas in the lab reports they are more complex, often using more specialist language than discharge and

progress notes, thus requiring external knowledge resources to resolve relations between terms. So domain knowledge helps with the latter, but tends to confound the former.

That said, the knowledge sources used – WordNet, UMLS/MetaMap, GSpell – were still insufficient to resolve some relations in these reports, although, arguably, some of the relations identified in the ground truth data as coreferential were not strictly so. For example, the ability to coreference carcinoma mentions that are linked to the formation of a mass, such as '*adenocarcinoma*' with '*exophytic mass*', or pairing histological studies such as '*chemical stains*' with '*MLH1*'. Similarly, in the discharge summaries/progress reports, the external domain knowledge sources were unable to resolve synonyms and metonyms such as '*confusion ... delirium*'; '*ecchymosis ... hematoma*'; '*carcinoma ... tumor*'; '*unable to ambulate ... bed bound*'; '*pins and needles from the knees ... neuropathic type pain*'. The limitations of these standard domain knowledge sources, despite their size and general comprehensiveness (see Chapter 6), in resolving these relations were noted by He[12], and the lack of improvement on system performance from the inclusion of external domain knowledge resources is consistent with previous studies on concept extraction from clinical notes[37].

Nevertheless, with or without the external domain resources, the system performs significantly better over the baseline overall. However, one reason for the lack of significant precision improvements over the baseline for the ODIE corpus may be explained by the difference in make up of the two data sets. In the ground truth ODIE data, around 80% of all mentions are in a coreference chain, but a typical ODIE document contains only about 6 such chains. So a baseline coreference of simply chaining, in each document, all mentions of the same class, has a reasonable chance that the selections made will be reasonably precise (mean F_1 over the Mayo and Pittsburgh data = 65%). However, for the i2b2/VA data, only $\approx$ 50% of mentions are in a coreference chain, yet there are on average about 16 chains per document, so the baseline method should perform less well (mean F_1 over the Beth Israel, Partners and Pittsburgh data = 52%).

For the system output, individual errors will have a greater impact on overall accuracy in documents with fewer anaphoric relations than in those with many relations. This

was typically the case with the ODIE corpus (on average 36.2 relations per document vs. 70.5 for the i2b2/VA corpus), which also had a higher mean chain length (5.7 vs. 4.3). These may partially explain the overall weaker ODIE results in comparison to those for the i2b2/VA corpus, although further work is needed to analyse performance in relation to coreference chain length in more detail. In addition, it has been suggested that some coreference evaluation metrics favour longer coreference chains[8].

In terms of individual mention classes, the system performed well at coreferencing `Person` mentions across all document types. Arguably, correct and precise identification of coreferential personal pronouns and name strings is more important in this context than for other mention types, where the goal might be to flag potentially linked events for review by clinicians as to the precise nature of the relationship. As noted in Table 8.14, it may be useful to link test, procedure and symptom/disorder events that are not strictly coreferential (but still relate to the same experiencer, i.e. the patient) in an event chain; for example, linking previous surgical procedures to a current surgical procedure for the same condition.

Overall, the results suggest that this approach developed here provides greatly increased coreference resolution performance in comparison to that reported for general-purpose tools (where F_1 ranges from 0% to 35%)[10]. In evaluating the performance of these tools[6][9], Hinote et al.[10] used the same corpora and coreference-specific evaluation metrics as our system, so the comparison is a reasonable one.

With some qualifications, the approach presented here also appears to offer an improvement over a number of previously reported clinical coreference systems[11][12][8]. Romauch[11] used a corpus of clinical guideline documents and did not detail the evaluation metrics used, so results may not be directly comparable. He[12] used a small corpus of 47 discharge summaries that may be similar to those in the i2b2/VA corpus, and reported scores from the B^3 and MUC metrics used in the current study, so comparison with the current results seems reasonable. Zheng et al.[8] reported results on a subset of the ODIE corpus used here and used the same evaluation metrics. However, their system performed end-to-end identification and coreference of clinical terms, whereas our system

(as with [10][11][12]) performs coreference only on existing mentions. Zheng et al. estimated that errors in term recognition accounted for 20% of system errors; it may be that extending this system to provide end-to-end evaluation would lead to a similar reduction in performance.

System results were submitted to the 2011 i2b2/VA Natural Language Processing Challenge for Clinical Records[13], where it ranked overall 7th out of 28 submissions to the 'coreference only' tracks. Precision against the i2b2/VA corpus was equal to that of the top-performing systems; for a full comparison, see Uzuner et al.[13][2]. Had system results without domain knowledge for the i2b2/VA corpus been submitted, the ranking would have been in the top 5. But more importantly, perhaps, this system appears to perform at least as well as human annotators — results for the ODIE corpus are comparable to the mean inter-annotator agreement (IAA) across the Mayo and UPMC datasets reported[14] of 75.4%, compared to 79.2% for this system. For the i2b2/VA corpus, a mean IAA across the Beth Israel, Partners Healthcare and UPMC datasets of 73.8% was reported (Appendix II of Uzuner et al.[13]) compared to 87.5% for this system.

This system does not impose a limit on the distance between coreferents. In contrast, Zheng et al.[8] imposed a 10 sentence window, as a sample of the training data suggested that a larger limit led to an unacceptable reduction in precision. However, they found that this limit was the most frequent source of recall error, as coreference relations can often span large distances, for example, between the History of Present Illness and Final Diagnosis sections at opposite ends of the document. Therefore, further work could involve examining the effect of varying the distance limit between mentions on the precision and recall of our system.

The acyclic, forward linked-list approach to coreference chain generation ensures that a given mention only participates in a single coreference chain. By cloning antecedent features to the anaphor and the use of a double-linked list, the narrative direction of the relationship is preserved, while all transitive coreference relationships can be extracted by traversing the linked list.

[2]Results presented in this chapter for the ODIE corpus differ from those reported in [13] as system output errors identified after submission have since been fixed

Although the results suggest that the patterns demonstrated here are reasonably generalizable across the 977 documents that came from a wide variety of sources, counter-examples can doubtless be found. Further work should investigate performance on clinical notes from other healthcare centre to determine the generalisability of this approach.

8.7. From coreference resolution to identifying processes of care

By way of a recap of the example given in Section 8.3.8, the extract below shows the annotations generated by the system, where the prepositions and verb groups identified by the ANNIE VG Chunker (see Chapter 5 and Figure 8.3) are highlighted in bold:

> [Mr WWWWW]$_{\text{Person,patient,male,singular}}$ [**is**]$_{\text{VG}}$ a [58 y/o]$_{\text{Age}}$ [gentleman]$_{\text{GenderIdentifier}}$ [*who*]$_{\text{Person,patient,male,singular}}$ [**was admitted**]$_{\text{VG}}$ ... **by** [Dr FFFF]$_{\text{Person,clinician,singular}}$. ... [He]$_{\text{Person,patient,male,singular}}$ [**was assessed**]$_{\text{VG}}$ **by** [Dr GGGGG]$_{\text{Person,clinician,singular}}$ ... [She]$_{\text{Person,clinician,female,singular}}$ [**has referred**]$_{\text{VG}}$ [WWW]$_{\text{Person,patient,male,singular}}$ **to** [the orthopedics team]$_{\text{Person,clinican,plural}}$; [he]$_{\text{Person,patient,male,singular}}$ [**will be followed**]$_{\text{VG}}$ **up by** [them]$_{\text{Person,clinician,plural}}$.

From this, the system generates the following `Person` coreference chains:

> P: (Mr WWWWW → who → He → WWW → he);
> C1: (Dr GGGGG → She);
> C2: (the orthopedics team → them);
> C3: the singleton (Dr FFFF).

where each protagonist has the following properties:

> patient(P)
> male(P)
> age(P, 58)
> clinician(C1)
> clinician(C2)

female(C2)

clinician(C3)

where P = 'Mr WWWWW', C1 = 'Dr FFFF', C2 = 'Dr GGGGG', C3 = 'the orthopedic team'. In combination with the verb phrases and prepositional phrases attached to each protagonist, these coreference chains allow us to generate the following, ordered narrative chains:

admitted(C1, P)

assessed(C2, P)

referred(C2, P)

follow_up(C3, P)

From this, we can create a narrative schema, along the lines proposed by Chambers and Jurafsky[7], as shown in Figure 8.8.

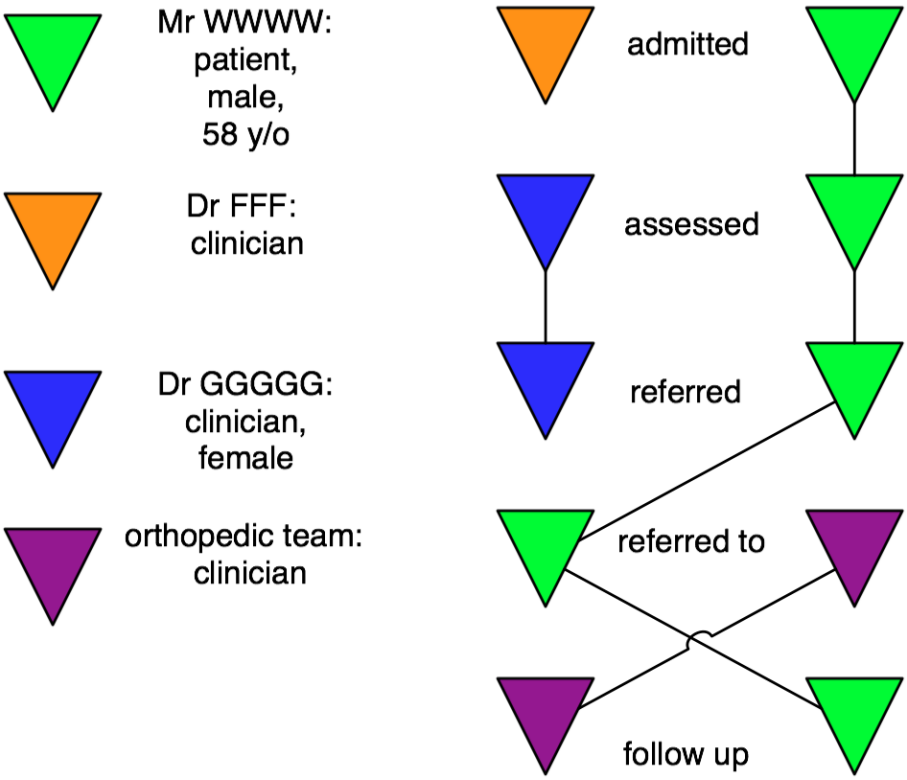


Figure 8.8.: Narrative schema for the text in Section 8.3.8

Obviously this is a simple example; in the general case, determining the correct entities to which to attach predicates (verb groups, prepositional phrases) is not straightforward and is difficult to generalise. However, such a schema could provide a useful, visual summary of the processes of care described in a patient's clinical notes. Chambers and Jurafsky[7] derived schema for narrative event chains from newswire text using unsupervised clustering techniques. In this work, we have used lexical patterns and external domain knowledge resources to generate coreference chains that follow the order of the narrative. Future work could investigate combining this work with the unsupervised learning of Chambers and Jurafsky to generate complete narrative schema from clinical texts.

8.8. Summary

The recent review by Zheng et al.[8] called on research into the portability of general coreference resolution methods to the clinical domain. We have combined these methods with additional patterns to address weaknesses in the general approaches when applied to clinical notes, namely integration of external domain knowledge, dealing with name deidentification/anonymisation, spelling errors and inconsistencies, use of abbreviations, and wide scope of pronominal resolution.

This approach augments generic methods (based on headword and pronoun-matching rules using gender, role, number and recency agreement) with specific approaches for clinical narratives: report section segmentation; abbreviation expansion; consideration of clinical relationships such as family history, quantitative, spatial, temporal, and anatomical contexts, and consideration of protagonists and their associated actions. Perhaps most importantly, we found that identifying synonym and hypernym relations via inclusion of external domain knowledge is most useful when processing lab reports, but is less useful in discharge summaries, where more straightforward pronoun resolution and context-sensitive string matching seem to be sufficient.

In the following chapter, we discuss the findings of the research presented in this and the preceding chapters, and review the aims and objectives stated in Chapter 1. The research

findings are discussed in terms of contributions to knowledge, limitations, and relation to other recent research in this area. Finally, we discuss options for future research.

9. Discussion and conclusion

9.1. Introduction

In the previous chapter (Chapter 8), a text processing pipeline for resolving pronominal and nominal coreference relations in a wide variety of clinical notes was developed and evaluated. The pipeline integrated the components developed and evaluated in Chapters 5–7, which were used for identifying the temporal, spatial and anatomical contexts of terms. Chapter 8 also described new components for pronoun classification, protagonist identification (patient, clinician, family relation), synonym matching (via WordNet), and coreference chain generation.

In this final chapter, we discuss the findings of the previous chapters in a wider context. One common thread that links the results for all the modules in the framework is that, with the exception of the method developed in Chapter 4 for identifying ontology terms, they generally performed with higher precision (0.77–0.98) than recall (0.62–0.97). This is consistent with other pattern generalisations, such as the Hearst patterns for identifying hypernym–hyponym relationships (see Chapter 6), although the patterns presented here still also perform with generally good recall. Achieving high precision in preference to recall may be desirable, as fewer pieces of correct information may be preferable to provision of larger amounts of information, through which the clinician needs to sift to discard the information not relevant. In the component developed for abbreviation expansion (see Chapter 7), a method for adjusting precision and recall dynamically was provided, so these trade-offs could be balanced. Parasuraman and Manzey[1] suggest that the cutoff value for the reliability (precision) of an automated system is around 70%: below this, the need to correct the information provided by the system outweighs the benefits of automation.

9.2. Review of aims, objectives, hypotheses and contributions to knowledge

At the start of this thesis, the stated the aim was to build on existing methods for automated identification and classification of clinical concepts and processes of care from heterogenous clinical texts, in order to facilitate the knowledge formalisation process for process-oriented clinical decision support systems. We have partially met this aim. A number of lexico-syntactic patterns and external knowledge integration components have been developed, and have been evaluated on clinical discharge summaries and progress notes, lab reports, MedLine abstracts, and clinical guideline documents. New techniques and tools for text segmentation (Chapter 5), concept identification and classification via semantic decomposition of ontologies (Chapter 6) and expansion and coreference of abbreviations (Chapter 7) have been developed and evaluated, and have been shown to offer improved performance over previous methods and systems. Recently, a number of these components have now been adopted as core modules within the suite of tools within GATE that provide support for biomedical text processing[1]. Since the work in Chapter 5 was completed, some of the ideas proposed in that chapter have been incorporated into the 2012 version of MetaMap. A `-composite_phrases` option has been introduced that identifies prepositional phrases of a user-defined length, and attempts to match these to precoordinated terms in the UMLS.

In building on existing work on concept identification and classification, a small amount of new work has been presented on word sense disambiguation (Chapters 7 and 8) and identification of negation and possibility (Chapter 5), but comprehensive coverage of these topics would need to be the subject of more in-depth, future research.

In terms of applying the techniques and tools to extracting knowledge from clinical guidelines to facilitate the knowledge formalisation process, only a formative evaluation and illustrative examples have been achieved (see Chapter 5). This is an area that would benefit from further development and evaluation against existing models derived from specific clinical guidelines, although apart from the models developed for the Protocure

[1] http://gate.ac.uk/sale/tao/splitch16.html#sec:domain-creole:biomed

project[2] or for the *OpenClinical* Guideline Modelling Methods Comparison[3], the availability of such corpora appear to be lacking. In the current work, the clinical guideline documents from the BMJ and NICE (see Chapter 4) were largely used as a training corpus from which to develop the phrase chunking patterns described in Chapter 5, prior to the availability of the clinical records from i2b2 on which the majority of this work has been evaluated.

In terms of identifying processes of care, a method for resolving coreference relations in clinical notes, and of generating coreference chains using progressively pruned linked lists, has been developed and evaluated (Chapter 8). We have shown how the approach provides a means to extract linked narrative events from clinical notes, although, as with the clinical guideline knowledge extraction, only illustrative examples have been demonstrated. However, in terms of coreference performance, the results have been independently evaluated, and performance exceeds that of general purpose tools, and is close to that of other recently reported, state-of-the art systems[2]. We have also demonstrated how incorporation of external knowledge resources improves coreference resolution performance for lab reports, but not for discharge summaries or progress notes. The difference in domain knowledge requirements between processing pathology reports and discharge summaries was also recently discussed by Rink et al.[3]; however, here we have quantified the effect.

9.2.1. Review of objectives

Chapter 2 provided an overview of knowledge representation for process-oriented clinical decision support systems (Objective #1). It did not go into detail about specific formalisms, for this, the reader is recommended to read reviews by Isern[4], Mulyar[5] and others, relevant chapters in Greenes[6], and formalism-specific syntax documentation (e.g. [7]).

Chapter 3 reviewed the current challenges in implementing process-oriented clinical systems and developed a conceptual implementation framework (Objectives #2 and #3). The framework showed how a clinical process model consists of a medical knowledge represen-

[2]http://www.protocure.org/old/resources-publications.html#ref-protocols
[3]http://www.openclinical.org/gmmcomparison.html

tation comprising clinical concepts and process knowledge mapped to EHR data items, combined with a localised knowledge representation comprising organisational workflow, local goals and temporal constraints. It used a novel methodology (thematic analysis and principle component analysis) to extract and relate themes from a large body of literature. However, the method and proposed model has not yet been validated by other researchers.

Chapter 3 also identified that, despite advances in clinical workflow, guideline and pathway modelling and architectures, a core knowledge acquisition problem remains, in terms of extracting and formalising the structured knowledge items required by these models, both from guidelines and the free text of the EHR. Chapters 4 and 5 identified that the proposed framework would sit in the upper left-hand corner of the conceptual model developed in Chapter 3 (Figure 3.5 on page 76). That is, the part of the model that utilises the integration of ontologies and pattern templates to identify and map concepts and processes (Objective #3). The framework also sits within the information quality axis in the DeLone–McClean[8] (D&M) model for information systems evaluation (see Chapter 4).

Chapters 4 and 5 identified that, despite advances in supervised machine learning techniques, there was still scope for development of explicit, lexico-syntactic patterns and rules for the creation of general-purpose, interoperable clinical information extraction components, which can be configured into pipelines for a range of tasks without requiring 'glue code'. These chapters identified current research challenges of clinical, quantitative, temporal and process concept identification and formalisation, spelling correction, abbreviation expansion, negation and coreference resolution (Objective #4):

Chapters 5, 6, 7 and 8 described the development of framework components that addressed the challenges listed above, and evaluated their performance against a number of publicly available, 'gold standard' research corpora (Objective #5). However, in these chapters we have provided only a technical evaluation of performance in terms of precision and recall against corpora manually annotated by domain experts. Further evaluation would need to see the framework integrated with an existing system and appraised by clinicians in terms of overall utility for real-world tasks. Some suggestions are given in

Section 9.3 below.

Overall then, within the scope set out in Chapter 1, the research objectives have been met for the corpora of clinical notes and MedLine abstracts used in the evaluation. However, further work remains in terms of evaluation against clinical guideline data.

9.2.2. Review of research hypotheses

In Chapter 1 two research hypotheses were stated: 1) that complex information extraction tasks could be assembled from self-contained components; and 2) that these components could be created from external knowledge sources, lexico-syntactic patterns and regular expressions. For the second hypothesis, Chapter 6 has described in detail and evaluated a method using regular expressions over domain ontology lexemes to create concept recognisers; Chapters 5 and 8 have provided numerous examples of lexico-syntactic patterns; and Chapter 7 has also demonstrated the use of regular expressions for abbreviation expansion.

For the first hypothesis, while some of the individual tasks addressed in Chapters 5–6 might be considered quite basic and fundamental, Chapter 8 suggests that the task of coreference resolution and narrative chain creation is quite complex. Moreover, Chapter 8 demonstrates how these individual, basic components can be assembled in a pipeline process to address this more complex task. One limitation of pipeline processes, however, where the output of one component is used as the input to another component, is that errors can propagate through the system. For example, if all components make use of the output of a part-of-speech tagger, then errors made by the tagger may result in a later contextual feature being missed, leading to term or relation misclassification. In Table 8.14 of Chapter 8, we noted an instance where domain knowledge, added earlier in the pipeline, led to an erroneous addition to a coreference chain, which led to other mentions being incorrectly added to the tail of the chain.

As noted in Chapter 4, the use of lexico-syntactic patterns and regular expressions over morphemes is not a new idea. The contribution to knowledge here is the development of expressions that generalise over clinical guidelines and patient notes, and the implementation of these patterns in the JAPE formalism. A limitation, however, of this approach for

identifying and classifying actual or potential ontology terms is that it does not provide the concept identifiers (e.g UMLS CUI) of the terms recognised, only the general classification. Lookup of the candidate terms against the UMLS using some best-match method would still be required to map the candidate term to a concept in the Metathesaurus.

If the evaluation results presented in these chapters can be considered satisfactory – and the performance comparison with existing tools and recent work suggests that this might be a fair assessment – then these research hypotheses, within the scope of this thesis, have been confirmed. As noted earlier, however, what is important is the clinical utility of information extracted by a system, rather than how closely a system can match withheld test data. In the quantitative evaluation of each the framework components, analysing the discrepancies between system output and the 'gold standard' annotations revealed, perhaps surprisingly, a variety of errors and inconsistencies in these corpora (see error analyses and discussions in Chapters 6, 7 and 8).

Similarly, errors were identified in one of the core components of the UMLS, the Foundational Model of Anatomy (see Chapter 6). Although the number of errors were small in comparison to the size of the ontology ($\approx 0.1\%$), each error is also replicated in the UMLS, either resulting in concepts that can only be identified via the incorrect spelling, or duplicate concepts with different UMLS identifiers[9]. Hopefully, these errors will be corrected in future releases of the FMA and UMLS.

These findings do raise questions about the validity of comparing the performance of different systems against these (and other) corpora. If the corpora themselves contain errors, even if only a few percent of the total, then simply chasing higher precision and recall scores by attempting to match the manual annotations as closely as possible – including their inconsistencies and omissions – may be a questionable exercise. In what way is a system that more closely matches such a corpus, by a few percent in comparison to a previous system, 'better' than that previous system? Conversely, can a system evaluated against the corrected version of the corpus be directly compared against one evaluated against the earlier, inconsistent version?

As noted in Chapter 8, one of the benefits of the deterministic, pattern-based approach

used in this research is that it allows such errors and inconsistencies in the evaluation corpora to be identified. Ultimately, though, systems that identify concepts and relations in clinical texts need to be evaluated in terms of their utility in assisting in clinical decision making. It is likely to be more important that a system identifies terms or relationships of potential interest or relevance to the clinician, than whether it has classified a text string as `Disease` rather than `Symptom`, or classified a relationship between two phrases as coreferential rather than historic–current or part–whole. In the following section, we discuss possible avenues for future research to address these points.

9.3. Further work

In this work, we have tackled the problem of extracting, from free text, processes of care that unfold over time by addressing constituent problems of identifying clinical events, their protagonists, temporal contexts and coreferential relations. In our system's internal representation of coreference chains, the direction of the relationship and the ordering of mentions is preserved. However, by definition, the coreference relationship is an identity relationship: it is transitive and symmetric, and the metrics for evaluating coreference performance, as used in the evaluation here, make use of this assumption. Moreover, the identity relationship implies temporal independence, i.e. identity at any time. In strict coreference, the ordering of the terms in the text – and thus the ordering of mentions in each chain and the order of the coreference chains themselves – is unimportant. This assumption has led to recent developments of semi-supervised, clustering methods for coreference resolution.

However, we have argued that, in clinical notes, the direction of the coreference relationship and the ordering of coreference chains may be important for identifying the chains of clinical reasoning, even if the events in the narrative are not described in chronological order. By making the relationship directional, in $A \rightarrow B \rightarrow C$ we allow the possibility that A may be directly coreferential to C but may need to be mediated by B; that is, $A \rightarrow C$ may not hold at any time, but only when $A \rightarrow B$. For example, from a pathology report in the ODIE evaluation corpus:

> [Invasive adenocarcinoma, grade 3 (of 4)]$_{\mathrm{A}}$, forming [a 3 cm mass]$_{\mathrm{B}}$. [The tumor]$_{\mathrm{C}}$ invades into and through the muscularis propria

terms A, B and C ('*Invasive adenocarcinoma, grade 3 (of 4)*', '*a 3 cm mass*', *the tumor*) have been marked as coreferent. But the carcinoma leads to the formation of the mass, not vice versa: the $A \rightarrow B$ relation is not symmetric as the two are not strictly interchangeable. Equally, the $A \rightarrow C$ relationship ('*Invasive adenocarcinoma*' $\rightarrow$ '*the tumor*') is mediated by knowledge of the $A \rightarrow B$ relation.

In another example

> The patient is [unable to ambulate]$_{\mathrm{A}}$ ... The patient is [bed bound]$_{\mathrm{B}}$ ...

terms A and B were marked as coreferent, but B arises as a result of A; there is an implicit directional relationship rather than an identity relationship. These examples may seem like splitting hairs, but, as we have argued in Chapter 8, the ordering provides information about the clinical reasoning process, so it makes sense to preserve it.

Similarly, noting an injury on admission as '*the patient's head wound laceration*' with later discussions about '*her scalp laceration*' – both expressions refer to the same general injury finding, but the latter is more specific, and occurs later in the narrative, so the implication is that investigation of the general injury led to the more specific finding. Maintaining the ordering and direction of the relationship in the system's internal representation preserves this reasoning process. Manning[10] argued that it has been difficult to obtain value from coreference resolution in real-world applications. It may be that ignoring the direction of the narrative, and modelling coreference relations as bags of equivalent terms, might be one reason for this.

We have argued that preserving this narrative order is important for using the coreference relations to generate narrative event chains. Future work should combine this with the work done on temporal concept and relation identification to extract *temporally* ordered, rather than narratively ordered, processes of care. In fact, such work is one of the tasks in the i2b2 2012 NLP Shared Task[4], of which some of the work described in

[4] https://www.i2b2.org/NLP/TemporalRelations/Call.php

Chapter 5 was part. This future work could make use of the temporal dependency structures described recently by Kolomiyets et al.[11], to identify TimeML `TLINK` relations (see Chapter 5). As noted in Chapters 5 and 8, in terms of parsing and formalising temporal expressions and relations in free text, more work is needed on more identifying the temporal context of events, linking of relative dates back to correct event antecedent date, and better distinguishing of historical events from those in the current episode. One approach might be to store all events and their dates as tuples, for example '*transferred to ICU on 06/23/2012*' $\rightarrow (transfer, ICU, 06/23/2012)$.

Since the recent availability of libraries such as SUTime[12] and HeidelTime[13], the need to develop and maintain a temporal expression parser for the framework developed in this thesis (Chapter 5) is probably now unnecessary. It may be better to integrate an external library such as HeidelTime into the framework than attempt to replicate it, although as noted in Chapter 5, the output of such libraries may not be optimal for clinical text and may still require post-processing.

In this work, we have provided only a technical evaluation of the framework developed. Integration of the framework within a larger system would require a more comprehensive evaluation involving other axes of the DeLone–McLean evaluation model (see Chapter 4) such as user satisfaction, individual and organisational impact. Kaplan [14] noted that in addition to consideration of the effect of a system on clinical and organisational performance, the concept of individual and organisational 'fit' (loosely corresponding to D&M's axis of 'user satisfaction') needs to be considered. 'Fit' can be considered according to various axes: such as local clinical workflow (see Chapters 2 and 3), or differing cultural values and goals between the developer and clinician. For example, identification of opportunities for decision support might represent, at two extremes, either an interesting computer science problem or a potential undermining of the art of clinical practice[14]. The concept of 'fit' might also be considered as part of the organisational change management process that forms part of the conceptual model developed in Chapter 3.

Although the concept dates back to the work of Chu et al.[15], recent tools that visualise care processes as interactive 'care maps', such as the *Map of Medicine*[16] and NICE

Pathways, may provide one approach to achieving organisational and clinical 'fit', possibly with a low barrier to entry in terms of delivering process knowledge from clinical guidelines direct to the clinician at the point of care. However, such visual information would need to be linked to patient data in some way. One of Sittig et al.'s 'grand challenges'[17] (see Chapter 1) was the need to provide 'at a glance' summaries of patient data. Potentially, the tools developed here could be extended in future work to provide visual summaries of temporally ordered care processes and events in the patient record.

Similarly, further research might consider how to extend the framework presented here to extract Map of Medicine-style visual algorithms or flowcharts directly from the text of clinical guidelines, although perhaps this may be too ambitious given that decision points and sequential vs parallel processes are usually implicit rather than explicitly represented in the guideline text[18]. However, previous research suggests that such higher-level, process information can be extracted with a level of precision (0.87)[19] that exceeds the Parasuraman and Manzey threshold (see Section 9.1).

Identifying fine-grained treatment information in guidelines would need to account for underspecified statements (underlined in the following examples), such as

'Avoid the use of highly intensive management strategies'

or

'initiate appropriate interventions'

and qualitative terms that need to be mapped to numeric values or ranges, such as

'The moderate use of alcohol may increase HDL-cholesterol'

or

'If blood pressure remains uncontrolled on adequate doses of three drugs'

While automated quantification and formalisation of such statements is probably unrealistic, there would be a role for NLP techniques to at least identify and classify these vague statements, either using a machine-learning or pattern based approach, perhaps in

a similar way to the detection of uncertainty via the presence of 'hedging' terms[20]. As we suggest in Chapter 5, the tools developed here have the potential to help automate the process of translating semi-structured or well-specified guideline statements to a more formalised representation.

Potentially, then, future work might make use of the components developed here to generate annotated clinical guideline text and historical patient notes for use as a CDSS knowledge base, following the work recently described by Wagholikar et al.[21]. In that study, the free text of Papanicolaou (Pap) reports was analysed by an NLP system to identify mentions of glandular epithelial cell abnormalities and squamous cell carcinoma findings. This was used to generate patient-specific cervical cancer screening recommendations. In a set of 74 test cases, 66 of the system recommendations matched those of a physician; on review, 7 of the discrepancies were deemed to be the result of physician error[21].

Alternatively, the tools and techniques developed here can be used for research purposes. The South London and Maudsley Hospital's (SLaM) Biomedical Research Centre (BRC) are currently using the GATE framework within their Case Register Interactive Search (CRIS) tool[22]. They have developed a number of information extraction rules using the JAPE formalism, as used in this work, to extract smoking status, Mini Mental State Exam results, and drug dosages, in order to identify potential participants in clinical trials. Their focus is on favouring higher precision over recall[22], as with the work presented here.

Applying information extraction and natural language processing techniques to the free text of clinical notes is not the only way of extracting process information from the patient record. *Process mining*[23] makes use of the time-stamped activity logs of workflow systems (or any system that logs user, time, and activity performed, such as an EHR) to extract common patterns of system interactions and their temporal relations, which can be used to generate a process map of sequential, parallel and branching activities (see Chapter 2). Process mining has recently been applied to EHR data to extract processes of care as recorded by the EHR for comparison with a defined care pathway[24][25], although some researchers have found the models produced by process mining algorithms

did not reflect clinical reality[26]. At present, no studies have combined IE/NLP on the unstructured text of clinical notes with process mining of EHR event logs, and this would be an interesting area for future research.

9.4. Conclusion

A conceptual model for the implementation of process-oriented health information systems has been developed. Such systems make use of a knowledge base derived from workflow models and clinical guidelines, and are integrated with an electronic health record via mapping of concepts to shared domain resources such as ontologies. The task of formalising and mapping knowledge items in guidelines and the EHR can be done manually, or it can be assisted with information extraction tools. However, the majority of clinical guideline information remains in text form, and much of the useful clinical information residing in the EHR is in the free text fields of progress notes and lab reports. Natural language processing techniques provide a means to add structure to these unstructured clinical texts.

Lexico-syntactic patterns, features, and domain knowledge resources for a tackling variety of information extraction tasks in the clinical domain have been developed and evaluated. Although a number of methods and software artefacts have been developed in this research, the main focus has been on identification and classification of clinical terms, and resolution of coreferential and anaphoric relations in clinical text. Generation of coreference chains provides a means to extract linked narrative events and processes of care from patient notes. We have shown that coreference performance in discharge summaries and progress notes is largely dependent on correct identification of protagonist chains (patient, clinician, family relation), pronominal resolution, and string matching that takes account of experiencer, temporal, spatial, and anatomical context; whereas for lab reports external domain knowledge is additionally required.

Where available, results have been compared against existing systems for solving these tasks and have been found to improve on them, or approach the performance of recently reported, state-of-the-art systems. The software artefacts have been made available as

open-source components within the General Architecture for Text Engineering framework.

Appendices

A. Principal component analysis matrix association scores

The table below shows final set of the 25 challenge theme variables derived from thematic analysis of papers reviewed in Chapter 3. As discussed in that chapter, the association between themes was explored using the Galaxy and Matrix views within the RefViz software. RefViz identified 10 clusters within the groupings of 5 themes assigned to each of the 108 papers. The weight of each theme within each cluster indicates the strength of association between the theme and the cluster, on a scale from -1 (strongest negative association) through 0 (no association) to +1 (strongest positive association). .

Table A.1.: RefViz Matrix view showing weighting of variables in each cluster

	Cluster ID									
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Clinical implementation	-0.13	0.38	-0.13	-0.25	0.43	-0.25	0	0.33	-0.17	0.04
Clinician attitude	0	0	0	0	0	0	0	0.67	0	-0.04
Complexity	-0.07	-0.13	0	0.75	-0.14	0	-0.13	0	0.08	0
Data mapping	-0.27	0	0.09	-0.25	-0.29	-0.25	0.38	-0.33	-0.25	0.21
Discrepancy	0.07	0	-0.04	0	0.14	0	0	0	-0.08	0.04
Exception handling	0.13	0	-0.04	0	0	0.25	0	0	0	-0.04
Execution	-0.13	0.75	-0.09	0.5	-0.14	0	0.13	0	0	-0.13
Expressivity	0	0	0	1	-0.14	0	0.13	0	-0.08	-0.08
Flexibility and adaptability	0.4	-0.13	-0.04	0	-0.14	0	0	0	-0.08	-0.04
Goal modelling	-0.07	0	-0.04	0	0	0.75	0.13	0	-0.08	0
Guideline translation	-0.27	-0.25	0.74	-0.25	-0.14	-0.25	-0.13	-0.33	0.08	-0.25
Information/rule extraction	-0.07	0	0.09	0	0	0	-0.13	0	0	0
Localisation	-0.13	0.38	0.17	-0.25	-0.14	-0.25	0	0	0	-0.08
Maintenance	-0.13	-0.13	-0.04	0	0.14	0	0.13	0	0.33	-0.08

Continued on next page

Table A.1 – continued from previous page

	Cluster ID									
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Model validation	-0.13	-0.13	-0.13	0	-0.14	0	0.5	0	0.58	-0.13
Model verification	-0.13	-0.13	-0.13	-0.25	-0.14	0	0.38	0	0.75	-0.17
Organisational change	-0.07	0	0	0	0.29	0	-0.13	0.67	-0.08	-0.04
Organisational modelling	0.73	0	-0.13	-0.25	-0.14	-0.25	-0.13	0	-0.17	-0.08
Process modelling	0.47	0	0.09	-0.5	-0.43	-0.5	-0.25	-0.33	-0.5	0.33
Reporting, querying and visualisation	-0.07	-0.13	-0.09	0	0.29	0	0	0	-0.08	0.08
Separation of concerns	-0.07	0	-0.04	0	0	0.25	0.13	0	-0.08	0.13
System architecture	-0.07	-0.13	-0.09	0	-0.14	0	-0.13	0	-0.08	0.25
Temporal abstraction	-0.13	0.13	-0.09	0.5	-0.14	0.75	0.13	0	-0.08	-0.04
Tooling	-0.2	0	0.04	-0.25	0.43	-0.25	-0.25	-0.33	-0.08	0.17
UI and usability	-0.13	-0.13	-0.09	0	0.57	0	-0.13	0	0	0.13

B. Notes on system architectures and prototype implementations

The table below provides summary details of the 54 system architecture and prototype implementation studies identified in systematic review of Chapter 3.

Table B.1.: System architectures and prototype implementations

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Tierney (1995)[1]	Guideline-based order entry recommendations for heart failure	Standalone desktop application	Networked microcomputer workstations	No	Yes	Tight coupling (hardcoded into patient record system)	Electronic order entry forms with automated guideline recommendations	Block structured, procedural logic rules
Barnes (1995)[2]	Architecture for guideline recommendations system	Loosely coupled client-server, Web CGI interface, C++ libraries	Hardware independent	No	No	Low-level C++ interfaces (no structured text data exchange)	Guideline knowledge base, guideline explainers, Web interface	Object model
Fox (1997)[3]	Guideline-based decision support system	Standalone desktop application; PROforma CIG model	PC	No	No		Graphical GL editor, knowledge base, enactment engine	Formal task-network model (PROforma)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Chu (1998)[4]	Prototype electronic care pathway; guideline visualisation; visual process map	Standalone desktop application	PC	No	No	Activity-on-node visualisation forerunner of Map of Medicine; does not provide decision support	Visualisation and recording: ordering and charting layer, outcomes pathway layer, intervention layer	General task-network model
Henry (1998)[5]	Guideline-based patient notes templates for an existing EHR	Document templates for commercial EHR system	PC (Windows)	No	Yes		Template manager within WAVE EPR system	Block structured, procedural logic rules
Miller (1999)[6]	Tool for rule-based guideline verification to identify incomplete rule sets	Standalone desktop application (Hypercard)	Mac	No	No	Primarily provides guideline verification	Rule and constraint builder	Block structured, procedural logic rules

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Bindels (2000)[7]	Guideline-based test ordering system	Standalone desktop application (Delphi)	PC (Windows)	No	No	Delphi system with guidelines presented in HTML window	Graphical guideline editor, knowledge base, CDSS reminder/alerts, order entry	Block structured, procedural logic rules
Miller (2000)[8]	Guideline-based immunisation forecasting and recommendation system	Callable MUMPS module, remote Web CGI module	Hardware-independent	No	Yes		C/C++ callable module	Block structured, procedural logic rules

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Quaglini (2000)[9]	System architecture for guideline-based, clinical workflow system for acute ischemic stroke	Oracle Workflow engine, client-server Web application	Not given	Yes	No		Graphical guideline editor, patient-flow simulator, guideline knowledge base, organisational ontology, workflow management system, process monitor, alert/reminder notifier	Formal workflow model; formal task-network model (GUIDE)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Quaglini (2000)[10]	Computerised guideline for pressure ulcer prevention	Client-server, tightly coupled EHR integration, Java	Not given	No	Yes	Provides lists of daily tasks, not real-time workflow decision support. tightly coupled (EPR needed modification)	Graphical guideline editor, guideline inference engine, terminology server	Formal task-network model (GUIDE)
Dadam (2000)[11]	Architecture for a computerised clinical workflow system	Distributed, multi-server architecture, Java, workflow API	Hardware independent	Yes	No	Loose coupling via communication and service layers	Graphical process modelling tool, organisation modelling tool, workflow client, workflow server (execution layer, service layer, data access layer)	General workflow model; Block structured, procedural logic rules

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Chu (2001)[12]	Implementation of a computerised clinical pathway management system	Standalone desktop application	PC/Windows	No	No*	*Not clear from the system description. Plans to extend prototype into a multi-user, networked application	Order entry system; documentation and information management system; clinical pathway progress trending and monitoring; pathway variance alerts	Not specified

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
de Clercq (2001)[13]	Various decision support systems customised for different settings	Plug-in architecture; Web application	PC (Windows)	No	Yes	EPR integration not explained, other than use of 'standard communication protocols'	Knowledge editor (acquisition tool), domain ontology (classes), method library (functions, operations), guideline library, compiled knowledge base, datasource manager (EPR data), event monitor, execution scheduler, action manager (reminders, alerts, emails)	Formal task-network model (GASTON)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Mikulich (2001)[14]	Guideline-based emergency department electronic charting system	Standalone application	PC	No	No		Rule-based expert system modules	Block structured, procedural logic rules
Miller (2001)[15]	Test case generator tool for guideline-based rule sets	Standalone application (Lisp, GLIF)	Not given	No	No	Primarily provides guideline verification	Rule and constraint builder	Formal task-network model (GLIF)
Terenziani (2001)[16]	Architecture for guideline representation and execution	Modular, Web application, Java, Oracle	Hardware-independent	No	No	Tight coupling, DB-centric. Guideline verification/consistency checking, temporal constraint checks	Graphical guideline encoding tool, guideline execution engine	General task-network model

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Panzarasa (2002)[17]	Guideline-based workflow management system ('careflow') for stroke rehabilitation	Three-tier, agent-based, Oracle workflow engine	Not given	Yes	Yes	Tight coupling, database-centric	Organisational ontology, activity management layer, data management layer, communication layer	General task-network model; general workflow model
Malamateniou (2003)[18]	Intranet-based, inter-organisational workflow management system for radiological procedures	Distributed, multi-site Web-based workflow application; SOA, Web services, SOAP/HTTP, IBM's MQ Series Workflow, Oracle, Java	Distributed local application, Web and database servers	Yes	Yes	Describes organisational workflow - clinical decision support not apparent	Local and global authorisation server, centralised workflow management system server, distributed application servers, VMR	General workflow model; document model

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Maviglia (2003)[19]	Guideline-based decision support system for chronic diseases	Not given	Not given	No	Yes	Not clear if Web-based, client-server	Graphical guideline editor, guideline knowledge base, inference/execution engine, Notifier, Event monitor, messaging, EHR	Formal task-network model (GLIF)
Poulymenopoulou (2003)[20]	Intranet-based, inter-organisational workflow management system for emergency department	Distributed, multi-site Web-based workflow application; SOA, Web services, SOAP/HTTP, Oracle, IBM's MQ Series Workflow, Java	Distributed local application, Web and database servers	Yes	Yes		Local and global authorisation server, centralised workflow management system server, distributed application servers, VMR	General workflow model

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Barretto (2003)[21]	Architecture for guideline-based decision support and workflow in chronic disease management	Java, SOA, Web services, SOAP/HTTP, commercial middleware	Hardware independent	Yes	Yes	Loose coupling via SOA	Graphical workflow editor, workflow engine, guideline knowledge base (GLIF), guideline execution engine, messaging/notification middleware	Formal task-network model (GLIF)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Shahar (2004)[22]	Electronic guideline library	Modular, distributed, client-server Web application, Microsoft SQL Server	Windows Server	No	Yes		Web-based graphical guideline editor/encoding tool, Web-based guideline search, guideline knowledge base, Web-based vocabulary server, permissions/authorisation manager, guideline interpreter, task-specific reasoner, medical knowledge base, data visualisation tool	Formal task-network model (Asbru)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Wang (2004)[23]	Execution engine for guidelines encoded in GLIF3 format	Modular, three-tier loosely coupled client-server, Java	Hardware-independent	No	Yes*	*EHR interface available, but semantics need to be locally defined and implemented	Guideline execution engine, messaging/communication layer, execution tracing system	Formal task-network model (GLIF)
Anand (2004)[24]	Guideline-based decision support system for pediatric clinic	C#/.NET, Perl, Microsoft SQL Server, tightly coupled client-server	Windows (2003) server, integrated printer and scanner	No*	Yes	*workflow process hardcoded into system. Provides paper output	Arden MLM knowledge base, data dictionary, patient data store, Arden MLM rule processor, HL7 interface to EHR, printing/scanning module	Block structured, procedural logic rules (Arden)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Ciccarese (2004)[25]	Architecture for distributed guideline management system	Java, SOA, Web services, SOAP/HTTP	Hardware independent	Yes	Yes	Separation of concerns; loose coupling	VMR, graphical guideline editor/formalizer; guideline repository; inference engine; reporting system	General workflow model; Formal task-network model (GUIDE)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Ciccarese (2005)[26]	Integrated guideline management and workflow management system	Java Web service application; SOA, Web services, SOAP/HTTP	Hardware independent	Yes	Yes	Separation of concerns; loose coupling	Graphical guideline editor, knowledge base (central and localised guideline templates), guideline enactment engine, messaging and integration interface, reporting, VMR	Formal workflow model; Formal task-network model (GUIDE)
Colombet (2005)[27]	Web-based decision support system for preventive medicine	PHP/Javascript Web application, Visio, Excel	Not given	No	No	No consideration of temporal constraints	Graphical guideline editor, knowledge base, execution engine	Block structured, procedural logic rules

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Michalowski (2005)[28]	Mobile, guideline-based decision support system for emergency triage	Extended, tightly coupled client-server, thick client (local database, ES shell)	Mobile device, wireless network	No*	Yes	*hardcoded workflow in the triage process. Fits clinical workflow via point-of-care access	Domain ontology, knowledge base, ES shell (solver, executor), HL7 communication layer, interface engine, client-server synchronisation subsystem	Block structured, procedural logic rules

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Seroussi (2005)[29]	Guideline-based decision support for hypertension management	Standalone Web application (interactive guideline decision tree browser)	Hardware-independent	No	No	Doesn't provide automated decision support, just allows clinician to navigate decision tree and input values for parameters	Guideline knowledge base, guideline visualisation browser	Block structured, procedural logic rules
Aigner (2006)[30]	User-interface prototype for guideline visualisation	MVC, Java, standalone desktop application	Not given	No	No	Provides temporal 'lifeline' view of clinical process	Graph visualizer (JGraph), view manager, event monitor	Formal task-network model (Asbru)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Burkle (2006)[31]	Prototype workflow model of clinical pathway for lumbar nerve root compression syndrome	Adonis workflow engine, web browser, Orbis EPR	Hospital-wide network of 2000 workstations	Yes	Yes*	*Complete pathway not implemented (only preadmission phase and discharge)	Visual workflow editor, application for diagnosis of LNRCS, application for activity task checklist generation, report generation	General workflow model
Hayward-Rowse (2006)[32]	Conversion of paper care pathway documentation to electronic forms	Microsoft Access; standalone desktop application	PC/Windows	No	No	No decision support or workflow, simply an electronic representation of existing paper forms	Care plan, risk assessment, patient medication record, manually created alerts	Not specified

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Panzarasa (2006)[33]	Guideline-based workflow management system ('careflow') for stroke rehabilitation	Oracle Workflow engine, middleware, tightly coupled, database-centric	Not given	Yes	Yes	Tight coupling, database-centric	Middleware EPR-workflow integration layer, event monitor	General task-network model; general workflow model
Vesely (2006)[34]	Automated guideline-based reminder system	Standalone interactive GLIF browser	Hardware-independent	No	No*	*EHR integration described in general terms but not implemented	Graphical guideline decision tree browser, guideline execution	Formal task-network model (GLIF)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Wakamiya (2006)[35]	Electronic management system for paper-based clinical pathways	Tightly coupled client-server, desktop client application, Visual C++, Filemaker, Excel	Windows NT server and Windows XP client terminals of an order entry system in 100BASE-T/10BASE-T Ethernet network	No	No		Patient registration, CP administration, CP task checklist editor, variance recording	Not specified
Kaiser (2007)[36]	Automated guideline formalisation application	Java, standalone desktop application	Not given	No	No	Unclear how automated formalisation can capture implicit knowledge	Information extraction pipeline for template generation	Formal task-network model (Asbru)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Lenz (2007)[37]	Guideline-based electronic clinical pathway	CASE tool for generating document-based modules for commercial EHR system (Orbis®/OpenMedical)	PC (Windows)	Yes*	Yes	*Not full workflow application, but 'workflow enabled forms'.	Graphical form generator and process flow tool, HL7 compliant interface engine	General workflow model
Leonardi (2007)[38]	Diabetes management workflow system	YAWL, agent-based, Web-service application	Not given	Yes	No*	Organisational ontology, serviceflow, Virtual Healthcare Organisation (like VMR but organisation side); contract, tight coupling. *Not clear	Communication layer, service layer, organisational units (workflow) layer, contract (what, when, who) layer	Formal workflow model (YAWL)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Young (2007)[39]	Guideline-based decision support system	Modular, distributed, loosely coupled client-server Web application, SOA, Web services	Not given	No	Yes	Loose coupling via Web service interfaces	Patient data access server, guideline library server (see Shahar (2004))	Formal task-network model (Asbru)
Sartipi (2007)[40]	Decision support architecture for Canada Infoway service bus	SOA, enterprise service bus, HL7 CDA	Not given	No*	No*	*Not yet implemented: conceptual architecture	Guideline-based workflow service, data mining service, healthcare network visualisation service,	General workflow model

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Verlaenen (2007)[41]	A decision support architecture for clinical workflow, pathways and guidelines	Client-server, Java Swing client, JBoss server, MySQL/Hibernate database, Drools and jRules rules engine; PROforma CIG model	Hardware independent	Yes	No*	* Via VMR and locally defined EHR-VMR mappings, but only workflow execution implemented here	Graphical workflow editor, VMR, clinical knowledge base, event monitor, workflow manager, execution environment, system integration layer	Formal task-network model (PROforma)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Alexandrou (2008)[42]	A semantic workflow management system to support the dynamic adaptation of clinical pathways	OWL ontology, SWRL, Java, Jess rules engine, ActiveBPEL workflow	Hardware independent	Yes	No		Semantic layer (knowledge base, SWRL rule base), adaptation layer (rule engine, clinical pathway creation and update manager), clinical pathway layer (process repository, workflow engine)	Semantic Web

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Casteleiro (2008)[43]	Simulation framework for generating guideline recommendations from online clinical data	Web service application; SOA, Web services, OWL-S, SWRL	Hardware-independent	No	No	Separation of concerns; loose coupling, HL7 RIM and CDA. Temporal constraints not considered.	Patient identification service, clinical information service, guideline recommendation service	Semantic Web
Dang (2008)[44]	Web based healthcare workflow application	Web service application; SOA, Web services, OWL-S, WSDL-S	Hardware-independent	Yes	No*	*Not explicitly mentioned.	Knowledge editor (Protege), knowledge base (Jena), Wf engine, service composer, task manager,	Semantic Web; general workflow model

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Goud (2008)[45]	Active (infobutton) guideline-based decision support system for cardiac rehabilitation, integrated with EHR	Tightly coupled client-server	Not given	No	Yes	Tightly coupled	Knowledge base, execution engine, CDSS communication layer	Formal task-network model (GASTON)
Patkar (2008)[46]	Implementation of breast cancer triple assessment care pathway	Standalone Web application, PROforma CIG model	Not given	No	No		Graphical workflow editor and guideline editing toolkit, guideline knowledge base, execution engine	Formal task-network model (PROforma)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Seyfang (2008)[47]	Guideline execution engine for high-frequency clinical domains	Not given	Not given	No	No*	*Simulated execution traces presented; only the interpreter has been developed so far	Guideline compiler, execution manager, execution tracing component	Formal task-network model (Asbru)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Allart (2009)[48]	ICU real-time bedside monitoring system integrated with medical sensors	Integrated, distributed hardware/software solution; modular (storage, acquisition, computation, display); tightly coupled client-server; Web-based (HTTP); thick client	Dedicated, secured network; distributed data clusters; biomedical sensor equipment; desktop PCs	No	Yes	Bespoke architecture; tight coupling	Knowledge base, messaging/-communication, modules layer: event monitoring, hardware integration (drivers), computing/inference, display and visualisation	Formal task-network model (Think! network)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Argüello (2009)[49]	Simulation framework for generating guideline recommendations from online clinical data	Web service application; SOA, Web services, OWL-S, SWRL	Hardware-independent	No	No	Separation of concerns; loose coupling, HL7 RIM and CDA	HL7 CDA patient data repository, guideline OWL knowledge base, SWRL rule base, reasoning engine, execution tracing system	Semantic Web

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Laleci (2009)[50]	SAPHIRE guideline execution, bedside monitoring and alerting prototype integrated with medical sensors	Web service application; SOA, Web services, OWL-S; IHE XDS	Biomedical sensor equipment integrated with software	Yes	Yes	VMR, HL7 CDA. 'IHE XDS exposes the semantics of clinical documents through document metadata', separations of concerns - WF integration handled by OWL-S	IHE XDS Repository, UDDI service registry, alerting/reminder system	Formal task-network model (GLIF); Semantic Web
Peleg (2009)[51]	Guideline-based decision support system for diabetic foot care	Standalone client (Java), GLEE (guideline execution engine for GLIF)	Hardware-independent	No	Yes		Guideline knowledge base, guideline execution engine, user interface	Formal task-network model (GLIF)

Continued on next page

Table B.1 – continued from previous page

Publication	Purpose	Software architecture	Hardware architecture	Wf support	EHR integration	Notes	Components	Model type
Daniyal (2009)[52]	Web-based care pathway for prostate cancer management	Finite state machine (FSM) workflow Web application; OWL ontology	Hardware independent	Yes	No	Pathway as a high-level process map	Care pathway knowledge base, care pathway execution engine	Semantic Web
Eccher (2009)[53]	Architecture for a guideline-based decision support system for cancer care	Web application, SOA, Web services	Hardware independent	No	Yes	Not yet implemented, conceptual architecture, loosely coupled	EPR/VMR mapping layer, guideline knowledge base, guideline interpreter,	Formal task-network model (Asbru)
Tschopp (2009)[54]	CPOE and clinical pathway system	jBPM workflow, service-based components, Java	Hardware independent	Yes	Yes*	*Appears to be closely coupled to in-house EHR. Unclear if Web-based or bespoke client	Message-driven middleware	General workflow model

C. Examples of gold standard, 'ground truth' labelled data sets

The unlabelled data sets that formed the input to the information extraction pipeline, and the manually annotated versions (of the same documents) against which the outputs of the pipeline were evaluated, are provided by the corpus creators in a variety of formats. Agreement on a standard format – at least in the clinical domain – seems to be some way off. As a result, in many cases, each format required an import routine to be written to be able to import the data into the framework. As described in Chapters 4 to 8, the labels and their character offsets were imported into a separate annotation set against which the system-generated labels from the plain text input could be evaluated using GATE's corpus quality assurance component. Withheld test data from the i2b2 challenges (see Chapters 5 and 8) were provided as unlabelled, plain-text data; the output of the pipeline was evaluated against the labelled data using scripts provided by i2b2.

Examples of each of the corpora used are provided in the following sections.

C.1. Data set for concept identification (Chapter 5)

In the ODIE and i2b2 data sets, plain-text discharge summaries and laboratory reports are provided along with a separate 'concepts' file detailing each concept text boundary, the corresponding line and word offsets in the plain text file, and the concept label name, as follows:

```
c="fibroadenoma" 3:73 3:73||t="diseaseorsyndrome"
c="the procedure" 4:13 4:14||t="procedure"
```

```
c="stereotactic biopsy" 12:3 12:4||t="procedure"
c="benign breast parenchyma" 5:16 5:18||t="anatomicalsite"
c="stereotactic biopsy" 3:61 3:62||t="procedure"
c="breast malignancy" 3:65 3:66||t="diseaseorsyndrome"
c="mrs. jjjjjj" 3:53 3:54||t="people"
c="repeat left breast mammogram" 3:91 3:94||t="procedure"
c="she" 4:10 4:10||t="people"
c="breast stereotactic biopsy" 3:17 3:19||t="procedure"
c="six-month left breast mammogram" 6:2 6:5||t="procedure"
c="skin" 3:38 3:38||t="anatomicalsite"
c="i" 3:79 3:79||t="people"
c="skin" 5:20 5:20||t="anatomicalsite"
c="pain" 4:78 4:78||t="signorsymptom"
c="i" 4:52 4:52||t="people"
c="mrs. jjjjjj" 3:2 3:3||t="people"
c="i" 3:0 3:0||t="people"
c="clinical breast exam" 6:7 6:9||t="procedure"
c="the radiology department" 4:29 4:31||t="people"
c="microcalcifications" 5:26 5:26||t="diseaseorsyndrome"
c="her" 4:68 4:68||t="people"
c="microcalcifications" 3:44 3:44||t="diseaseorsyndrome"
c="the stereotactic biopsy" 3:10 3:12||t="procedure"
c="mrs. jjjjjj" 4:56 4:57||t="people"
c="benign breast parenchyma" 3:34 3:36||t="anatomicalsite"
c="the patient" 3:13 3:14||t="people"
c="painful" 4:17 4:17||t="signorsymptom"
c="mrs. jjjjjj" 3:82 3:83||t="people"
c="which" 3:74 3:74||t="none"
c="that appointment" 3:102 3:103||t="other"
```

```
c="skin" 12:20 12:20||t="anatomicalsite"
c="i" 4:65 4:65||t="people"
c="i" 3:50 3:50||t="people"
c="stereotactic biopsy" 5:3 5:4||t="procedure"
c="clinical breast examination" 3:96 3:98||t="procedure"
c="the stereotactic biopsy" 4:6 4:8||t="procedure"
c="we" 3:99 3:99||t="people"
c="she" 4:19 4:19||t="people"
c="benign breast parenchyma" 12:16 12:18||t="anatomicalsite"
c="mrs. jjjjjj" 4:0 4:1||t="people"
c="i" 4:24 4:24||t="people"
c="the biopsy" 3:68 3:69||t="procedure"
c="her" 3:105 3:105||t="people"
c="she" 3:85 3:85||t="people"
c="we" 4:38 4:38||t="people"
c="microcalcifications" 12:26 12:26||t="diseaseorsyndrome"
```

As shown in the example above, the string '*benign breast parenchyma*' occurring on line 5, words 16-18 of the plain-text report should be annotated by the system as `AnatomicalSite` to count as a match with the ground-truth concepts (where a word is determined in the plain-text input as a character string delimited by non-word character such as white space or punctuation). Below is the (anonymised) plain-text document that accompanies this concept file.

> I telephoned Mrs. JJJJJJ today to deliver the results of the stereotactic biopsy. The patient underwent left breast stereotactic biopsy at the 3 o'clock and 9 o'clock position. The pathology report demonstrates fragments of benign breast parenchyma and skin showing features of a fibroadenoma. Microcalcifications were present in the specimen. I have reassured Mrs. JJJJJJ that there was no evidence on stereotactic biopsy of a breast malignancy. Rather, the biopsy is consistent with fibroadenoma, which is a benign process. I have

advised Mrs. JJJJJJ that she should return in six-months for repeat left breast mammogram and clinical breast examination. We will schedule that appointment for her. Mrs. JJJJJJ was quite upset about the stereotactic biopsy procedure. She reports that the procedure was extremely painful and she was given minimal anesthesia. I will discuss this with the radiology department to see if there is anything we can do in the future to minimize the pain associated with this procedure. I have also provided Mrs. JJJJJJ with a prescription for Percocet, 10 tablets. I have advised her to take 1-2 tablets every six hours as-needed for pain. #1 Left breast stereotactic biopsy at the 3 o'clock and 9 o'clock position demonstrating fragments of benign breast parenchyma and skin showing features of a fibroadenoma. Microcalcifications were present in the specimen. #2 Advised six-month left breast mammogram and clinical breast exam

C.2. Data set for identification and expansion of abbreviations (Chapter 7)

As described in Chapter 7, the BioText 'yeast' corpus consisted of a plain text file of 1000 MedLine abstracts, in which abbreviations and their expansions are delimited by `<Short>` and `<Long>` tags, respectively. Matching `id` attribute values connect long form–short form pairs, as shown in the extract below.

```
The <Long id="120">yeast cadmium factor</Long> (<Short id="120">Ycf1p</Short>)
is a vacuolar <Long id="121">ATP binding cassette</Long>
(<Short id="121">ABC</Short>) transporter required for heavy metal
and drug detoxification. Cluster analysis shows that Ycf1p is strongly
related to the human <Long id="122">multidrug-associated protein</Long>
(<Short id="122">MRP1</Short>) and cystic fibrosis transmembrane
conductance regulator and therefore may serve as an excellent
model for the study of eukaryotic ABC transporter structure
```

```
and function. ...

Thirteen intragenic second-site suppressors were identified for
the  D777N mutation which affects the invariant Asp residue in
the Walker B motif of the <Long id="123">first nucleotide binding
domain</Long> (<Short id="123">NBD1</Short>). Two of the suppressor
mutations (V543I and F565L) are located in the <Long id="124">
first transmembrane domain</Long> (<Short id="124">TMD1</Short>),
nine (A1003V, A1021T, A1021V, N1027D, Q1107R, G1207D, G1207S, S1212L,
and W1225C) are found within TMD2, one (S674L) is in NBD1, and
another one (R1415G) is in NBD2, indicating either physical proximity
or functional interactions between NBD1 and the other three domains.
```

C.3. Data set for event and temporal relation identification (Chapter 5)

In contrast to the multiple input files for each report provided by the developers of the concept and coreference identification corpora (see Section C.1 above and Section C.4 below), a single input file for each report was provided for the event and temporal relation identification task. Here, the `EVENT` and `TIMEX3` tags point to concept boundary start and end character offsets, label (via the `type` attribute), and label feature values for modality, polarity, normalised value and modifier.

<ClinicalNarrativeTemporalAnnotation>
<TEXT><![CDATA[
Admission Date :
02/19/1993
Discharge Date :
02/25/1993
HISTORY OF PRESENT ILLNESS :

C. Examples of gold standard, 'ground truth' labelled data sets

This is an 72-year-old who presented with postmenopausal spotting and had an endometrial biopsy which was read at the Etearal Etsystems/
Hospital as showing grade I adenocarcinoma .
Accordingly she presents for operative therapy at this time .
HOSPITAL COURSE :
The patient was brought to the Operating Room on 2-19-91 where she had an exploratory laparotomy , TAH / BSO , and omental biopsy .
She had a normal abdominal exploration of a small uterus with superficial invasion on gross examination , normal ovaries .
Washings were sent .
A subfascial J-P was left .
The patient did well postoperatively and had a regular diet by the third postoperative day .
Subfascia drain was discontinued .
The patient 's postoperative Hct was 34 .
]]></TEXT>
<TAGS>
<EVENT id="E0" start="1" end="10" text="Admission" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E2" start="113" end="122" text="presented" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E3" start="128" end="151" text="postmenopausal spotting" modality="FACTUAL" polarity="POS" type="PROBLEM" />
<EVENT id="E4" start="160" end="181" text="an endometrial biopsy" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E5" start="192" end="196" text="read" modality="FACTUAL" polarity="POS" type="EVIDENTIAL" />
<EVENT id="E6" start="200" end="231" text="the Etearal Etsystems/ Hospital" modality="FACTUAL" polarity="POS" type="CLINICAL_DEPT" />

<EVENT id="E7" start="235" end="242" text="showing" modality="FACTUAL" polarity="POS" type="EVIDENTIAL" />
<EVENT id="E8" start="243" end="265" text="grade I adenocarcinoma" modality="FACTUAL" polarity="POS" type="PROBLEM" />
<EVENT id="E9" start="284" end="292" text="presents" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E1" start="29" end="38" text="Discharge" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E10" start="297" end="314" text="operative therapy" modality="FACTUAL" polarity="POS" type="TREATMENT" />
<EVENT id="E11" start="375" end="393" text="the Operating Room" modality="FACTUAL" polarity="POS" type="CLINICAL_DEPT" />
<EVENT id="E12" start="419" end="444" text="an exploratory laparotomy" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E13" start="447" end="456" text="TAH / BSO" modality="FACTUAL" polarity="POS" type="TREATMENT" />
<EVENT id="E14" start="463" end="477" text="omental biopsy" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E15" start="488" end="518" text="a normal abdominal exploration" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E25" start="497" end="518" text="abdominal exploration" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E17" start="542" end="562" text="superficial invasion" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E16" start="566" end="583" text="gross examination" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E18" start="603" end="611" text="Washings" modality="FACTUAL" polarity="POS" type="TEST" />
<EVENT id="E19" start="617" end="621" text="sent" modality="FACTUAL"

polarity="POS" type="OCCURRENCE" />
<EVENT id="E20" start="624" end="640" text="A subfascial J-P" modality="FACTUAL" polarity="POS" type="TREATMENT" />
<EVENT id="E21" start="668" end="672" text="well" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E22" start="697" end="711" text="a regular diet" modality="FACTUAL" polarity="POS" type="OCCURRENCE" />
<EVENT id="E23" start="745" end="760" text="Subfascia drain" modality="FACTUAL" polarity="POS" type="TREATMENT" />
<EVENT id="E24" start="814" end="817" text="Hct" modality="FACTUAL" polarity="POS" type="TEST" />
<TIMEX3 id="T0" start="18" end="28" text="02/19/1991" type="DATE" val="1991-02-19" mod="NA" />
<TIMEX3 id="T2" start="318" end="327" text="this time" type="DATE" val="1991-02-19" mod="NA" />
<TIMEX3 id="T3" start="397" end="404" text="2-19-91" type="DATE" val="1991-02-19" mod="NA" />
<TIMEX3 id="T1" start="46" end="56" text="02/25/1991" type="DATE" val="1991-02-25" mod="NA" />
<TIMEX3 id="T4" start="715" end="742" text="the third postoperative day" type="DATE" val="1997-02-22" mod="NA" />
</TAGS>
</ClinicalNarrativeTemporalAnnotation>

C.4. Data set for coreference resolution (Chapter 8)

Chains of co-referring terms from the concept boundary and classification ground-truth data (see Section C.1) were provided as a separate file for each clinical report. The 'chains' file for the example report and concepts file given in Section C.1 is shown in the extract

below.

```
c="I" 3:0 3:0||c="I" 3:50 3:50||c="I" 3:79 3:79||c="I" 4:24 4:24
||c="the radiology department" 4:29 4:31||c="we" 4:38 4:38
||c="we" 4:38 4:38||c="I" 4:52 4:52||c="I" 4:65 4:65||t="coref people"
c="painful" 4:17 4:17||c="pain" 4:78 4:78||t="coref signorsymptom"
c="repeat left breast mammogram" 3:91 3:94
||c="clinical breast examination" 3:96 3:98
||c="that appointment" 3:102 3:103
||c="six-month left breast mammogram" 6:2 6:5
||c="clinical breast exam" 6:7 6:9
||t="coref procedure"
c="skin" 3:38 3:38||c="skin" 5:20 5:20||c="skin" 12:20 12:20
||t="coref anatomicalsite"
c="Mrs. JJJJJJ" 3:2 3:3||c="The patient" 3:13 3:14
||c="Mrs. JJJJJJ" 3:53 3:54
||c="Mrs. JJJJJJ" 3:82 3:83||c="she" 3:85 3:85
||c="her" 3:105 3:105
||c="Mrs. JJJJJJ" 4:0 4:1||c="She" 4:10 4:10||c="she" 4:19 4:19
||c="Mrs. JJJJJJ" 4:56 4:57||c="her" 4:68 4:68||t="coref people"
c="fibroadenoma" 3:73 3:73||c="which" 3:74 3:74
||t="coref diseaseorsyndrome"
c="Microcalcifications" 3:44 3:44||c="Microcalcifications" 5:26 5:26
||c="Microcalcifications" 12:26 12:26||t="coref diseaseorsyndrome"
c="the stereotactic biopsy" 3:10 3:12
||c="breast stereotactic biopsy" 3:17 3:19
||c="stereotactic biopsy" 3:61 3:62||c="the biopsy" 3:68 3:69
||c="the stereotactic biopsy" 4:6 4:8||c="the procedure" 4:13 4:14
||c="stereotactic biopsy" 5:3 5:4
||c="stereotactic biopsy" 12:3 12:4||t="coref procedure"
```

```
c="benign breast parenchyma" 3:34 3:36
||c="benign breast parenchyma" 5:16 5:18
||c="benign breast parenchyma" 12:16 12:18||t="coref anatomicalsite"
```

As shown in this extract, concepts of '*repeat left breast mammogram*', '*clinical breast examination*', '*that appointment*', '*six-month left breast mammogram*' and '*clinical breast exam*', occurring at the corresponding line/word offsets in the plain text of the report, were considered by the human annotators to refer to the same `Procedure` event, so the system output should aim to match this.

References for Chapter 1

[1] P. D. Clayton and G. Hripcsak. Decision support in healthcare. *Int. J. Biomed. Comput.*, 39(1):59–66, Apr 1995.

[2] J. Fox, E. Black, I. Chronakis, R. Dunlop, V. Patkar, M. South, and R. Thomson. From guidelines to careflows: modelling and supporting complex clinical processes. *Stud Health Technol Inform.*, 139:44–62, 2008.

[3] M. Peleg and S. Tu. Decision support, knowledge representation and management in medicine. *Yearb Med Inform*, pages 72–80, 2006.

[4] Samson W. Tu, James R. Campbell, Julie Glasgow, Mark A. Nyman, Robert McClure, James McClay, Craig Parker, Karen M. Hrabak, David Berg, Tony Weida, James G. Mansfield, Mark A. Musen, and Robert M. Abarbanel. The sage guideline model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5):589–598, 2007. doi: DOI: 10.1197/jamia.M2399.

[5] Mor Peleg, Sagi Keren, and Yaron Denekamp. Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (kdom). *Journal of Biomedical Informatics*, 41(1):180–201, 2008. doi: DOI: 10.1016/j.jbi.2007.05.003.

[6] Leila Ahmadian, Mariette van Engen-Verheul, Ferishta Bakhshi-Raiez, Niels Peek, Ronald Cornet, and Nicolette F. de Keizer. The role of standardized data and terminological systems in computerized clinical decision support systems: Literature review and survey. *International Journal of Medical Informatics*, 80(2):81 – 93, 2011.

[7] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates. Grand challenges in clinical decision support. *J Biomed Inform*, 41(2):387–392, Apr 2008.

[8] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. doi: DOI: 10.1016/j.jbi.2009.08.007.

[9] Robert A. Greenes. A proposed strategy for overcoming inertia. In R. A. Greenes, editor, *Clinical Decision Support: The Road Ahead*, pages 541–80. Academic Press, Burlington, 2007. doi: DOI: 10.1016/B978-012369377-8/50003-9.

[10] Robert A. Greenes. Definition, scope and challenges. In R. A. Greenes, editor, *Clinical Decision Support: The Road Ahead*, pages 2–19. Academic Press, Burlington, 2007. doi: DOI: 10.1016/B978-012369377-8/50004-0.

[11] J. Fox, D. Glasspool, V. Patkar, M. Austin, L. Black, M. South, D. Robertson, and C. Vincent. Delivering clinical decision support services: there is nothing as practical as a good theory. *J Biomed Inform*, 43(5):831–843, Oct 2010.

[12] K. B. Wagholikar, K. L. Maclaughlin, M. R. Henry, R. A. Greenes, R. A. Hankey, H. Liu, and R. Chaudhry. Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc*, 19(5):833–839, Sep 2012.

[13] Arnold W. Pratt and Milos G. Pacak. Automated processing of medical english. In *Proceedings of the 1969 conference on Computational linguistics*, COLING '69, pages 1–23, Stroudsburg, PA, USA, 1969. Association for Computational Linguistics.

[14] Ö Uzuner, Y Juo, and P Szolovits. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–63, 2007.

[15] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[16] C. Friedman, O. P. Alderson, H. J. Austin, J. J. Cimino, and B. S. Johnson. A General Natural-Language Text Processor for Clinical Radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, March 1994.

[17] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[18] Leonard D'Avolio, Dina Demner-Fushman, and Wendy W. Chapman. An introduction to clinical natural language processing. In *AMIA 2011 Annual Symposium, NLM Staff Papers and Presentations*, 2011.

[19] Mark Stevenson, Yikun Guo, Abdulaziz Al Amri, and Robert Gaizauskas. Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 71–79, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[20] P. Gooch and A. Roudsari. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*, 18(6):738–748, 2011.

[21] P. Gooch and A. Roudsari. A tool for enhancing metamap performance when annotating clinical guideline documents with umls concepts. In *IDAMAP Workshop at the 13th Conference on Artificial Intelligence in Medicine (AIME'11)*, 2011.

[22] P. Gooch. Badrex: In situ expansion and coreference of biomedical abbreviations using dynamic regular expressions. Technical report, City University London, 2012.

[23] P. Gooch and A. Roudsari. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *Journal of Biomedical Informatics*, 2012. doi: DOI: 10.1016/j.jbi.2012.02.012.

References for Chapter 2

[1] Workflow Management Coalition. Wfmc-tc-1011 ver 3 terminology and glossary english. Technical report, 1999.

[2] F. Malamateniou and G. Vassilacopoulos. Developing a virtual patient record using xml and web-based workflow technologies. *International Journal of Medical Informatics*, 70(2-3):131–139, 2003. doi: DOI: 10.1016/S1386-5056(03)00039-X.

[3] Zahra Niazkhani, Habibollah Pirnejad, Marc Berg, and Jos Aarts. The impact of computerized provider order entry systems on inpatient clinical workflow: a literature review. *Journal of the American Medical Informatics Association*, 16(4):539–549, 2009. doi: DOI: 10.1197/jamia.M2419.

[4] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang. Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc*, 8(6):527–534, 2001.

[5] P. D. Clayton and G. Hripcsak. Decision support in healthcare. *Int. J. Biomed. Comput.*, 39(1):59–66, Apr 1995.

[6] D. W. Bates, G. J. Kuperman, S. Wang, T. Gandhi, A. Kittler, L. Volk, C. Spurr, R. Khorasani, M. Tanasijevic, and B. Middleton. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Inform Assoc*, 10(6):523–530, 2003.

[7] J. J. Cimino. Infobuttons: anticipatory passive decision support. In *AMIA Annu Symp Proc.*, pages 1203–4, 2008.

[8] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330:765, 2005.

[9] M.J. Field and K.N. Lohr. *Clinical Practice Guidelines: Directions for a New Program*. Institute of Medicine, Washington, DC, 1990.

[10] John Fox, Alyssa Alabassi, Vivek Patkar, Tony Rose, and Elizabeth Black. An ontological approach to modelling tasks and goals. *Computers in Biology and Medicine*, 36(7-8):837–856, 2006. doi: DOI: 10.1016/j.compbiomed.2005.04.011.

[11] European Pathways Association. Clinical / care pathways, 2007. Available from: http://www.e-p-a.org/000000979b08f9803/index.html (accessed 06 June 2012).

[12] K. Zander. Nursing case management: strategic management of cost and quality outcomes. *J Nurs Admin*, 18:23–30, 1988.

[13] Shunji Wakamiya and Kazunobu Yamauchi. What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*, 78(8):543–550, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.03.003.

[14] Harry Campbell, Rona Hotchkiss, Nicola Bradshaw, and Mary Porteous. Integrated care pathways. *BMJ*, 316(7125):133–137, 1998.

[15] T Rotter, L Kinsman, E James, A Machotta, H Gothe, J Willis, P Snow, and J. Kugler. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Technical report, Department of Public Health, Dresden Medical School, University of Dresden, 2010.

[16] Sally C. Brailsford. System dynamics: what's in it for healthcare simulation modellers? In S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, editors, *Proceedings of the 2008 Winter Simulation Conference*, 2008.

[17] Korina Katsaliaki, Sally Brailsford, David Browning, and Peter Knight. Mapping care

pathways for the elderly. *Journal of Health Organization and Management*, 19(1):57–72, 2008.

[18] Dawn Carnes, Jayne Gallagher, Sally Herne, Elaine Munday, Sara Ritchie, and Martin Underwood. Mapping care pathways and estimating the number and cost of musculoskeletal chronic pain patients to inform the development and implementation of a new service. *Primary Health Care Research and Development*, 9:241–247, 2008.

[19] OpenClinical. Clinical pathways: multidisciplinary plans of best clinical practice, 2003. Available from: http://www.openclinical.org/clinicalpathways.html (accessed 06 June 2012).

[20] Workflow Management Coalition. Process definition interface – xml process definition language. Technical report, 2008.

[21] A.-W. Scheer, O. Thomas, and O. Adam. Process modeling using event-driven process chains. In M. Dumas, W. M. P. van der Aalst, and A. H. M. ter Hofstede, editors, *Process-aware Information Systems : Bridging People and Software through Process Technology.*, page 119–145. Wiley, Hoboken, New Jersey, 2005.

[22] Arthur H. ter Hofstede and Wil M. van der Aalst. Yawl: yet another workflow language. *Information Systems*, 30(4):245–275, 2005.

[23] J.L. Peterson. *Petri Net Theory and the Modeling of Systems.* Prentice-Hall, 1981.

[24] W.M.P. van der Aalst. The application of petri nets to workflow management. *The Journal of Circuits, Systems and Computers*, 8(1):21–66, 1998.

[25] A Wombacher, P Fankhauser, and E Neuhold. Transforming bpel into annotated deterministic finite state automata for service discovery, 2004.

[26] E. Oren and A. Haller. Formal frameworks for workflow modelling. Technical report, DERI – Digital Enterprise Research Institute, National University of Ireland, Galway, 2005.

[27] Carlo Combi, Yuval Shahar, Ameen Abu-Hanna, Marco Beccuti, Alessio Bottrighi, Giuliana Franceschinis, Stefania Montani, and Paolo Terenziani. Modeling clinical guidelines through petri nets. In *Artificial Intelligence in Medicine*, volume 5651 of *Lecture Notes in Computer Science*, pages 61–70. Springer Berlin / Heidelberg, 2009.

[28] John Fox, Nicky Johns, Colin Lyons, Ali Rahmanzadeh, Richard Thomson, and Peter Wilson. Proforma: a general technology for clinical decision support systems. *Computer Methods and Programs in Biomedicine*, 54(1-2):59–67, 1997. doi: DOI: 10.1016/S0169-2607(97)00034-5.

[29] María Adela Grando, David W. Glasspool, and John Fox. Petri nets as a formalism for comparing expressiveness of workflow-based clinical guideline languages. In Will Aalst, John Mylopoulos, Norman M. Sadeh, Michael J. Shaw, and Clemens Szyperski, editors, *Business Process Management Workshops*, volume 17 of *Lecture Notes in Business Information Processing*, pages 348–360. Springer Berlin Heidelberg, 2009.

[30] Annette ten Teije, Mar Marcos, Michel Balser, Joyce van Croonenborg, Christoph Duelli, Frank van Harmelen, Peter Lucas, Silvia Miksch, Wolfgang Reif, Kitty Rosenbrand, and Andreas Seyfang. Improving medical protocols by formal methods. *Artificial Intelligence in Medicine*, 36(3):193–209, 2006. doi: DOI: 10.1016/j.artmed.2005.10.006.

[31] Alessio Bottrighi, Laura Giordano, Gianpaolo Molino, Stefania Montani, Paolo Terenziani, and Mauro Torchio. Adopting model checking techniques for clinical guidelines verification. *Artificial Intelligence in Medicine*, 48(1):1–19, 2009. doi: DOI: 10.1016/j.artmed.2009.09.003.

[32] W.M.P van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. Workflow patterns. *Distributed and Parallel Databases*, 14(3):5–51, 2003.

[33] W.M.P. van der Aalst. Three good reasons for using a petri-net-based workflow management system. In *Proceedings of the International Working Conference on Infor-*

mation and Process Integration in Enterprises (IPIC'96), pages 179–201, Cambridge, Massachusetts, 1996.

[34] F.S. Hillier and G.J. Lieberman. Project management with pert/cpm. In *Introduction to Operations Research*. McGraw-Hill, Boston, MA, 2010.

[35] H.L. Gantt. In *Work, Wages and Profit*. Hive Publishing Company, Easton, PA, 1974.

[36] S. Chu and B. Cesnik. Improving clinical pathway design: lessons learned from a computerised prototype. *International Journal of Medical Informatics*, 51(1):1–11, 1998.

[37] Yuval Shahar, Silvia Miksch, and Peter Johnson. The asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine*, 14(1-2):29–51, 1998. doi: DOI: 10.1016/S0933-3657(98)00015-3.

[38] RG Coyle. *System Dynamics Modelling. A practical approach*. Chapman and Hall, London, 1996.

[39] P. Checkland and J. Scholes. *Soft Systems Methodology in Action*. Wiley, Chichester, 1990.

[40] RL Flood and ER Carson. *Dealing with Complexity*. Plenum Press, New York, 1993.

[41] C. Rosse and J.V.L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36:478–500, 2003.

[42] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[43] Luca Anselma, Paolo Terenziani, Stefania Montani, and Alessio Bottrighi. Towards a comprehensive treatment of repetitions, periodicity and temporal constraints in clinical guidelines. *Artificial Intelligence in Medicine*, 38(2):171–195, 2006. doi: DOI: 10.1016/j.artmed.2006.03.007.

[44] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*, pages 28–34. AAAI Press, 2003.

[45] R. Goud, N. F. de Keizer, G. ter Riet, J. C. Wyatt, A. Hasman, I. M. Hellemans, and N. Peek. Effect of guideline based computerised decision support on decision making of multidisciplinary teams: cluster randomised trial in cardiac rehabilitation. *BMJ*, 338:b1440, 2009.

[46] Y. Shahar, O. Young, E. Shalom, M. Galperin, A. Mayaffit, R. Moskovitch, and A. Hessing. A framework for a distributed, hybrid, multiple-ontology clinical-guideline library, and automated guideline-support tools. *J Biomed Inform*, 37(5):325–44, 2004.

[47] S. Quaglini, M. Grandi, P. Baiardi, M. C. Mazzoleni, C. Fassino, G. Franchi, and S. Melino. A computerized guideline for pressure ulcer prevention. *Int J Med Inform*, 58-59:207–17, 2000.

[48] Mor Peleg, Lily A. Gutnik, Vincenza Snow, and Vimla L. Patel. Interpreting procedures from descriptive guidelines. *Journal of Biomedical Informatics*, 39(2):184–195, 2006. doi: DOI: 10.1016/j.jbi.2005.06.002.

[49] Vimla L. Patel, José F. Arocha, Melissa Diermeier, Robert A. Greenes, and Edward H. Shortliffe. Methods of cognitive analysis to support the design and evaluation of biomedical systems: the case of clinical practice guidelines. *Journal of Biomedical Informatics*, 34(1):52–66, 2001. doi: DOI: 10.1006/jbin.2001.1002.

[50] R. N. Shiffman, G. Michel, A. Essaihi, and E. Thornquist. Bridging the guideline implementation gap: a systematic, document-centered approach to guideline implementation. *J Am Med Inform Assoc*, 11(5):418–26, 2004.

[51] Rick Goud, Mariette van Engen-Verheul, Nicolette F. de Keizer, Roland Bal, Arie Hasman, Irene M. Hellemans, and Niels Peek. The effect of computerized decision support on barriers to guideline implementation: A qualitative study in outpatient

cardiac rehabilitation. *International Journal of Medical Informatics*, 79(6):430–7, 2010. doi: DOI: 10.1016/j.ijmedinf.2010.03.001.

[52] MD Cabana, CS Rand, NR Powe, AW Wu, MH Wilson, PA Abboud, and HR Rubin. Why don't physicians follow clinical practice guidelines? a framework for improvement. *JAMA*, 282(15):1458–65, 1999.

[53] M Butzlaff, HC Vollmar, B Floer, N Koneczny, J Isfort, and S Lange. Learning with computerized guidelines in general practice? a randomized controlled trial. *Family Practice*, 21(2):183–188, 2003.

[54] J. J. Stolte, J. Ash, and H. Chin. The dissemination of clinical practice guidelines over an intranet: an evaluation. *Proc AMIA Symp*, pages 960–4, 1999.

[55] W. M. Tierney, J. M. Overhage, M. D. Murray, L. E. Harris, X. H. Zhou, G. J. Eckert, F. E. Smith, N. Nienaber, C. J. McDonald, and F. D. Wolinsky. Effects of computerized guidelines for managing heart disease in primary care. *J Gen Intern Med*, 18(12):967–76, 2003.

[56] V. Sintchenko, E. Coiera, J. R. Iredell, and G. L. Gilbert. Comparative impact of guidelines, clinical data, and decision support on prescribing decisions: an interactive web experiment with simulated cases. *J Am Med Inform Assoc*, 11(1):71–7, 2004.

[57] D. L. Schriger, L. J. Baraff, K. Buller, M. A. Shendrikar, S. Nagda, E. J. Lin, V. J. Mikulich, and S. Cretin. Implementation of clinical guidelines via a computer charting system: effect on the care of febrile children less than three years of age. *J Am Med Inform Assoc*, 7(2):186–95, 2000.

[58] W. M. Tierney. Improving clinical decisions and outcomes with information: a review. *Int J Med Inform*, 62(1):1–9, 2001.

[59] Rick Goud, Arie Hasman, and Niels Peek. Development of a guideline-based decision support system with explanation facilities for outpatient therapy. *Computer Methods and Programs in Biomedicine*, 91(2):145–153, 2008. doi: DOI: 10.1016/j.cmpb.2008.03.006.

[60] W. M. Tierney, J. M. Overhage, M. D. Murray, L. E. Harris, X. H. Zhou, G. J. Eckert, F. E. Smith, N. Nienaber, C. J. McDonald, and F. D. Wolinsky. Can computer-generated evidence-based care suggestions enhance evidence-based management of asthma and chronic obstructive pulmonary disease? a randomized, controlled trial. *Health Serv Res*, 40(2):477–97, 2005.

[61] M. W. Jaspers, M. Smeulers, H. Vermeulen, and L. W. Peute. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *J Am Med Inform Assoc*, 18(3):327–334, May 2011.

[62] R. N. Shiffman, Y. Liaw, C. A. Brandt, and G. J. Corb. Computer-based guideline implementation systems: a systematic review of functionality and effectiveness. *J Am Med Inform Assoc*, 6(2):104–14, 1999.

[63] G. Damiani, L. Pinnarelli, S. C. Colosimo, R. Almiento, L. Sicuro, R. Galasso, L. Sommella, and W. Ricciardi. The effectiveness of computerized clinical guidelines in the process of care: a systematic review. *BMC Health Serv Res*, 10:2, 2010.

[64] David Isern and Antonio Moreno. Computer-based execution of clinical guidelines: A review. *International Journal of Medical Informatics*, 77(12):787–808, 2008. doi: DOI: 10.1016/j.ijmedinf.2008.05.010.

[65] M Peleg, S Tu, J Bury, P Ciccarese, J Fox, RA Greenes, R Hall, PD Johnson, N Jones, A Kumar, S Miksch, S Quaglini, A Seyfang, EH Shortliffe, and M. Stefanelli. Comparing computer-interpretable guideline models: a case-study approach. *J Am Med Inform Assoc*, 10(1):52–68, 2003.

[66] P. De Clercq, K. Kaiser, and A. Hasman. Computer-interpretable guideline formalisms. *Stud Health Technol Inform*, 139:22–43, 2008.

[67] SW Tu, PD Johnson, and MA Musen. A typology for modeling processes in clinical guidelines and protocols. Technical report, Stanford Medical Informatics, 2002.

[68] B. Seroussi, J. Bouaud, and G. Chatellier. Guideline-based modeling of therapeutic strategies in the special case of chronic diseases. *International Journal of Medical Informatics*, 74(2-4):89–99, 2005. doi: DOI: 10.1016/j.ijmedinf.2004.06.004.

[69] Rianne Bindels, Paul A. de Clercq, Ron A. G. Winkens, and Arie Hasman. A test ordering system with automated reminders for primary care based on practice guidelines. *International Journal of Medical Informatics*, 58-59:219–233, 2000. doi: DOI: 10.1016/S1386-5056(00)00089-7.

[70] Y. Shahar, S. Miksch, and P. Johnson. An intention-based language for representing clinical guidelines. *Proc AMIA Annu Fall Symp*, pages 592–6, 1996.

[71] S. Quaglini, M. Stefanelli, A. Cavallini, G. Micieli, C. Fassino, and C. Mossa. Guideline-based careflow systems. *Artificial Intelligence in Medicine*, 20(1):5–22, 2000. doi: DOI: 10.1016/S0933-3657(00)00050-6.

[72] P. Dadam, M. Reichert, and K. Kuhn. Clinical workflows - the killer application for process-oriented information systems? In *Proc. 4th Int'l Conf. on Business Information Systems (BIS '00)*, pages 36–59, Poznan, Poland, 2000.

[73] Aziz A. Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T. Zeng, Dongwen Wang, Vimla L. Patel, Robert A. Greenes, and Edward H. Shortliffe. Glif3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of Biomedical Informatics*, 37(3):147–161, 2004. doi: DOI: 10.1016/j.jbi.2004.04.002.

[74] Samson Tu, James Campbell, and Mark Musen. The structure of guideline recommendations: a synthesis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 679–683, 2003.

[75] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. doi: DOI: 10.1016/j.jbi.2009.08.007.

[76] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[77] G. Hripcsak. Arden syntax for medical logic modules. *MD Comput*, 8(2):76, 78, 1991.

[78] RN Shiffman, BT Karras, A Agrawal, R Chen, L Marenco, and S. Nath. Gem: a proposal for a more comprehensive guideline document model using xml. *J Am Med Inform Assoc*, 7(5):488–98, 2000.

[79] Margarita Sordo, Omolola Ogunyemi, Aziz A. Boxwala, and Robert A. Greenes. Gello: An object-oriented query and expression language for clinical decision support. In *AMIA Annu Symp Proc.*, page 1012, 2003.

[80] Samson W. Tu, James R. Campbell, Julie Glasgow, Mark A. Nyman, Robert McClure, James McClay, Craig Parker, Karen M. Hrabak, David Berg, Tony Weida, James G. Mansfield, Mark A. Musen, and Robert M. Abarbanel. The sage guideline model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5):589–598, 2007. doi: DOI: 10.1197/jamia.M2399.

[81] H. C. Karadimas, C. Chailloleau, F. Hemery, J. Simonnet, and E. Lepage. Arden/J: an architecture for MLM execution on the Java platform. *J Am Med Inform Assoc*, 9(4):359–368, 2002.

[82] Nataliya Mulyar, Wil M. P. van der Aalst, and Mor Peleg. A pattern-based analysis of clinical computer-interpretable guideline modeling languages. *Journal of the American Medical Informatics Association*, 14(6):781–787, 2007. doi: DOI: 10.1197/jamia.M2389.

[83] G. Schadow, D. C. Russler, and C. J. McDonald. Conceptual alignment of electronic health record data with guideline and workflow knowledge. *International Journal of Medical Informatics*, 64(2-3):259–274, 2001.

[84] P. Gooch and A. Roudsari. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*, 18(6):738–748, 2011.

[85] Gillian Hardstone, Mark Hartswood, Rob Procter, Alex Voss, and Gwyneth Rees. Supporting informality: team working and integrated care records. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 142–151, 2004.

[86] Jason J. Saleem, Alissa L. Russ, Connie F. Justice, Heather Hagg, Patricia R. Ebright, Peter A. Woodbridge, and Bradley N. Doebbeling. Exploring the persistence of paper with the electronic health record. *International Journal of Medical Informatics*, 78(9):618–628, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.04.001.

[87] Adela Grando, Mor Peleg, and David Glasspool. A goal-oriented framework for specifying clinical guidelines and handling medical errors. *Journal of Biomedical Informatics*, 43(2):287–99, 2010. doi: DOI: 10.1016/j.jbi.2009.11.006.

[88] T. Benson. Care pathways. Technical report, National Programme for Information Technology, 2005.

[89] K. de Luc and J. Todd. Introduction. In K. de Luc and J. Todd, editors, *e-Pathways: computers and the patient's journey through care*, pages 1–14. Radcliffe Medical Press, Oxford, 2003.

[90] R. Page and I. Herbert. Developing e-pathway standards. In K. de Luc and J. Todd, editors, *e-Pathways: computers and the patient's journey through care*, pages 155–182. Radcliffe Medical Press, Oxford, 2003.

[91] K. de Luc and J. Todd. A way forward? In K. de Luc and J. Todd, editors, *e-Pathways: computers and the patient's journey through care*, pages 184–198. Radcliffe Medical Press, Oxford, 2003.

[92] Mor Peleg, Aziz A. Boxwala, Samson Tu, Qing Zeng, Omolola Ogunyemi, Dongwen Wang, Vimla L. Patel, Robert A. Greenes, and Edward H. Shortliffe. The

intermed approach to sharable computer-interpretable guidelines: a review. *Journal of the American Medical Informatics Association*, 11(1):1–10, 2004. doi: DOI: 10.1197/jamia.M1399.

[93] J. Todd. A systems view of care pathways. In K. de Luc and J. Todd, editors, *e-Pathways: computers and the patient's journey through care*, pages 109–154. Radcliffe Medical Press, Oxford, 2003.

[94] Kim M. Unertl, Matthew B. Weinger, Kevin B. Johnson, and Nancy M. Lorenzi. Describing and modeling workflow and information flow in chronic disease care. *Journal of the American Medical Informatics Association*, 16(6):826–836, 2009. doi: DOI: 10.1197/jamia.M3000.

[95] Jiangbo Dang, Amir Hedayati, Ken Hampel, and Candemir Toklu. An ontological knowledge framework for adaptive medical workflow. *Journal of Biomedical Informatics*, 41(5):829–836, 2008. doi: DOI: 10.1016/j.jbi.2008.05.012.

[96] J. Fox, E. Black, I. Chronakis, R. Dunlop, V. Patkar, M. South, and R. Thomson. From guidelines to careflows: modelling and supporting complex clinical processes. *Stud Health Technol Inform.*, 139:44–62, 2008.

[97] S. Chu and B. Cesnik. Modelling computerised clinical pathways. *Stud Health Technol Inform*, 52 Pt 1:559–63, 1998.

[98] V. Patkar and J. Fox. Clinical guidelines and care pathways: a case study applying proforma decision support technology to the breast cancer care pathway. *Stud Health Technol Inform*, 139:233–42, 2008.

References for Chapter 3

[1] P. Gooch and A. Roudsari. Computerization of workflows, guidelines, and care pathways: a review of implementation challenges for process-oriented health information systems. *J Am Med Inform Assoc*, 18(6):738–748, 2011.

[2] M. Schneider. Clinical information systems: strategic imperatives driving idns forward. *Med Netw Strategy Rep*, 8(11):1, 8–10, 12, 1999.

[3] A. Kushniruk, E. Borycki, S. Kuwata, and J. Kannry. Predicting changes in workflow resulting from healthcare information systems: ensuring the safety of healthcare. *Healthc Q*, 9 Spec No:114–8, 2006.

[4] Jos Aarts, Joan Ash, and Marc Berg. Extending the understanding of computerized physician order entry: Implications for professional collaboration, workflow and quality of care. *International Journal of Medical Informatics*, 76(Supplement 1):S4–S13, 2007. doi: DOI: 10.1016/j.ijmedinf.2006.05.009.

[5] Xiping Song, Beatrice Hwong, Gilberto Matos, Arnold Rudorfer, Christopher Nelson, Minmin Han, and Andrei Girenkov. Understanding requirements for computer-aided healthcare workflows: experiences and challenges. *ICSE '06: Proceedings of the 28th international conference on Software engineering*, pages 930–934, 2006.

[6] T Greenhalgh and R Taylor. How to read a paper: Papers that go beyond numbers (qualitative research). *BMJ*, 315:740–743, 1997.

[7] DN Evans and A Pearson. Systematic reviews of qualitative research. *Clinical effectiveness in Nursing*, 5:111–119, 2001.

[8] M Dixon-Woods, S Bonas, A Booth, DR Jones, T Miller, AJ Sutton, RL Shaw, JA Smith, and B Young. How can systematic reviews incorporate qualitative research? a critical perspective. *Qualitative Research*, 6:27–44, 2006.

[9] M. Dixon-Woods and R. Fitzpatrick. Qualitative research in systematic reviews. *BMJ*, 323:765–6, 2001.

[10] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ*, 330:765, 2005.

[11] T Rotter, L Kinsman, E James, A Machotta, H Gothe, J Willis, P Snow, and J. Kugler. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Technical report, Department of Public Health, Dresden Medical School, University of Dresden, 2010.

[12] Martin Eccles, Elaine McColl, Nick Steen, Nikki Rousseau, Jeremy Grimshaw, David Parkin, and Ian Purves. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. *BMJ*, 325(7370):941–, 2002.

[13] Emmy Rood, Robert Jan Bosman, Johan Ids van der Spoel, Paul Taylor, and Durk Freark Zandstra. Use of a computerized guideline for glucose regulation in the intensive care unit improved both guideline adherence and glucose regulation. *Journal of the American Medical Informatics Association*, 12(2):172–180, 2005. doi: DOI: 10.1197/jamia.M1598.

[14] N. Burns. Standards for qualitative research. *Nursing Science Quarterly*, 2:4–52, 1989.

[15] U. Flick. Coding and categorizing. In *An Introduction to Qualitative Research*, pages 305–332. Sage, London, 4th edition, 2009.

[16] M.B. Miles and A.M. Huberman. Early steps in analysis. In *Qualitative Data Analysis: An expanded sourcebook*, pages 55–89. Sage, Thousand Oaks, CA, 1994.

[17] Nancy R. Glassman. Refviz 1.0.1. *J Med Libr Assoc*, 93(2):293–294, 2005.

[18] Thomson Reuters. Ris format specifications, 2001.

[19] I.T. Jolliffe. *Principal Component Analysis*. Springer, second edition, 2002.

[20] Robert A. Greenes. Definition, scope and challenges. In R. A. Greenes, editor, *Clinical Decision Support: The Road Ahead*, pages 2–19. Academic Press, Burlington, 2007. doi: DOI: 10.1016/B978-012369377-8/50004-0.

[21] G. Schadow, D. C. Russler, and C. J. McDonald. Conceptual alignment of electronic health record data with guideline and workflow knowledge. *International Journal of Medical Informatics*, 64(2-3):259–274, 2001.

[22] S. Chu and B. Cesnik. Improving clinical pathway design: lessons learned from a computerised prototype. *International Journal of Medical Informatics*, 51(1):1–11, 1998.

[23] Erez Shalom, Yuval Shahar, Meirav Taieb-Maimon, Guy Bar, Avi Yarkoni, Ohad Young, Susana B. Martins, Laszlo Vaszar, Mary K. Goldstein, Yair Liel, Akiva Leibowitz, Tal Marom, and Eitan Lunenfeld. A quantitative assessment of a methodology for collaborative specification and evaluation of clinical guidelines. *Journal of Biomedical Informatics*, 41(6):889–903, 2008. doi: DOI: 10.1016/j.jbi.2008.04.009.

[24] Wolfgang Aigner and Silvia Miksch. Carevis: Integrated visualization of computerized protocols and temporal patient data. *Artificial Intelligence in Medicine*, 37(3):203–218, 2006. doi: DOI: 10.1016/j.artmed.2006.04.002.

[25] Aziz A. Boxwala, Mor Peleg, Samson Tu, Omolola Ogunyemi, Qing T. Zeng, Dongwen Wang, Vimla L. Patel, Robert A. Greenes, and Edward H. Shortliffe. Glif3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of Biomedical Informatics*, 37(3):147–161, 2004. doi: DOI: 10.1016/j.jbi.2004.04.002.

[26] J. M. Brokel, M. G. Shaw, and C. Nicholson. Expert clinical rules automate steps in delivering evidence-based care in the electronic health record. *Comput Inform Nurs*, 24(4):196–205, 2006.

[27] P. Ciccarese, E. Caffi, S. Quaglini, and M. Stefanelli. Architectures and tools for innovative health information systems: the guide project. *Int J Med Inform*, 74(7-8):553–62, 2005.

[28] Carlo Combi, Matteo Gozzi, Barbara Oliboni, Jose M. Juarez, and Roque Marin. Temporal similarity measures for querying clinical workflows. *Artificial Intelligence in Medicine*, 46(1):37–54, 2009. doi: DOI: 10.1016/j.artmed.2008.07.013.

[29] John Fox, Nicky Johns, Colin Lyons, Ali Rahmanzadeh, Richard Thomson, and Peter Wilson. Proforma: a general technology for clinical decision support systems. *Computer Methods and Programs in Biomedicine*, 54(1-2):59–67, 1997. doi: DOI: 10.1016/S0169-2607(97)00034-5.

[30] L. Hayward-Rowse and T. Whittle. A pilot project to design, implement and evaluate an electronic integrated care pathway. *J Nurs Manag*, 14(7):564–71, 2006.

[31] S. Hoelzer, R. Schweiger, and J. Dudeck. Representation of practice guidelines with xml–modeling with xml schema. *Methods Inf Med*, 41(4):305–12, 2002.

[32] Agnieszka Latoszek-Berendsen, Jan Talmon, Paul de Clercq, and Arie Hasman. With good intentions. *International Journal of Medical Informatics*, 76(Supplement 3):S440–S446, 2007. doi: DOI: 10.1016/j.ijmedinf.2007.05.012.

[33] R. Lenz, R. Blaser, M. Beyer, O. Heger, C. Biber, M. Bäumlein, and M. Schnabel. It support for clinical pathways: lessons learned. *International Journal of Medical Informatics*, 76(Supplement 3):S397–S402, 2007. doi: DOI: 10.1016/j.ijmedinf.2007.04.012.

[34] W. Michalowski, R. Slowinski, S. Wilk, K. J. Farion, J. Pike, and S. Rubin. Design and development of a mobile system for supporting emergency triage. *Methods Inf Med*, 44(1):14–24, 2005.

[35] D. Alexandrou, F. Xenikoudakis, and G. Mentzas. Adaptive clinical pathways with semantic web rules. In *Proceedings of the First International Conference on Health Informatics, HEALTHINF 2008*, volume 2, pages 140–147, Funchal, Madeira, Portugal, 2008. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

[36] S. A. Barretto, J. Warren, A. Goodchild, L. Bird, S. Heard, and M. Stumptner. Linking guidelines to electronic health record design for improved chronic disease management. *AMIA Annu Symp Proc*, pages 66–70, 2003.

[37] K. Bernstein, M. Bruun-Rasmussen, and S. Vingtoft. A method for specification of structured clinical content in electronic health records. *Stud Health Technol Inform*, 124:515–21, 2006.

[38] R. Chen, P. Georgii-Hemming, and H. Ahlfeldt. Representing a chemotherapy guideline using openehr and rules. *Stud Health Technol Inform*, 150:653–7, 2009.

[39] P. Ciccarese, E. Caffi, L. Boiocchi, S. Quaglini, and M. Stefanelli. A guideline management system. *Stud Health Technol Inform*, 107(Pt 1):28–32, 2004.

[40] V. Ebrahiminia, C. Duclos, M. E. Toussi, C. Riou, R. Cohen, and A. Venot. Representing the patient's therapeutic history in medical records and in guideline recommendations for chronic diseases using a unique model. *Stud Health Technol Inform*, 116:101–6, 2005.

[41] C. Eccher, A. Seyfang, A. Ferro, and S. Miksch. Embedding oncologic protocols into the provision of care: the oncocure project. *Stud Health Technol Inform*, 150:663–7, 2009.

[42] V. Patkar and J. Fox. Clinical guidelines and care pathways: a case study applying proforma decision support technology to the breast cancer care pathway. *Stud Health Technol Inform*, 139:233–42, 2008.

[43] Sartipi, Kamran, H. Yarmand Mohammad, and G. Down Douglas. Mined-knowledge and decision support services in electronic health. In *Proceedings of the Interna-*

tional Workshop on Systems Development in SOA Environments, pages 10–10, 2007. http://dx.doi.org/10.1109/SDSOA.2007.9.

[44] M. Tschopp, M. Despond, D. Grauser, J. C. Staub, and C. Lovis. Computer-based physician order entry: implementation of clinical pathways. *Stud Health Technol Inform*, 150:673–7, 2009.

[45] Mor Peleg, Lily A. Gutnik, Vincenza Snow, and Vimla L. Patel. Interpreting procedures from descriptive guidelines. *Journal of Biomedical Informatics*, 39(2):184–195, 2006. doi: DOI: 10.1016/j.jbi.2005.06.002.

[46] Mor Peleg, Sagi Keren, and Yaron Denekamp. Mapping computerized clinical guidelines to electronic medical records: Knowledge-data ontological mapper (kdom). *Journal of Biomedical Informatics*, 41(1):180–201, 2008. doi: DOI: 10.1016/j.jbi.2007.05.003.

[47] Mor Peleg, Aviv Shachak, Dongwen Wang, and Eddy Karnieli. Using multi-perspective methodologies to study users' interactions with the prototype front end of a guideline-based decision support system for diabetic foot care. *International Journal of Medical Informatics*, 78(7):482–493, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.02.008.

[48] Radu Serban, Annette ten Teije, Frank van Harmelen, Mar Marcos, and Cristina Polo-Conde. Extraction and use of linguistic patterns for modelling medical guidelines. *Artificial Intelligence in Medicine*, 39(2):137–149, 2007. doi: DOI: 10.1016/j.artmed.2006.07.012.

[49] B. Seroussi, J. Bouaud, and G. Chatellier. Guideline-based modeling of therapeutic strategies in the special case of chronic diseases. *International Journal of Medical Informatics*, 74(2-4):89–99, 2005. doi: DOI: 10.1016/j.ijmedinf.2004.06.004.

[50] Y. Shahar, O. Young, E. Shalom, M. Galperin, A. Mayaffit, R. Moskovitch, and A. Hessing. A framework for a distributed, hybrid, multiple-ontology clinical-

guideline library, and automated guideline-support tools. *J Biomed Inform*, 37(5):325–44, 2004.

[51] W. M. Tierney, J. M. Overhage, M. D. Murray, L. E. Harris, X. H. Zhou, G. J. Eckert, F. E. Smith, N. Nienaber, C. J. McDonald, and F. D. Wolinsky. Can computer-generated evidence-based care suggestions enhance evidence-based management of asthma and chronic obstructive pulmonary disease? a randomized, controlled trial. *Health Serv Res*, 40(2):477–97, 2005.

[52] Samson W. Tu, James R. Campbell, Julie Glasgow, Mark A. Nyman, Robert McClure, James McClay, Craig Parker, Karen M. Hrabak, David Berg, Tony Weida, James G. Mansfield, Mark A. Musen, and Robert M. Abarbanel. The sage guideline model: achievements and overview. *Journal of the American Medical Informatics Association*, 14(5):589–598, 2007. doi: DOI: 10.1197/jamia.M2399.

[53] O. Young, Y. Shahar, Y. Liel, E. Lunenfeld, G. Bar, E. Shalom, S. B. Martins, L. T. Vaszar, T. Marom, and M. K. Goldstein. Runtime application of hybrid-asbru clinical guidelines. *J Biomed Inform*, 40(5):507–26, 2007.

[54] C. J. Green, P. Fortin, M. Maclure, A. Macgregor, and S. Robinson. Information system support as a critical success factor for chronic disease management: Necessary but not sufficient. *Int J Med Inform*, 75(12):818–28, 2006.

[55] S. B. Henry, K. Douglas, G. Galzagorry, A. Lahey, and W. L. Holzemer. A template-based approach to support utilization of clinical practice guidelines within an electronic health record. *J Am Med Inform Assoc*, 5(3):237–44, 1998.

[56] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[57] Vladislav J. Mikulich, Yi-Ching A. Liu, Jennifer Steinfeldt, and David L. Schriger. Implementation of clinical guidelines through an electronic medical record: physician

usage, satisfaction and assessment. *International Journal of Medical Informatics*, 63(3):169–178, 2001. doi: DOI: 10.1016/S1386-5056(01)00177-0.

[58] P. L. Miller and S. J. Frawley. Trade-offs in producing patient-specific recommendations from a computer-based clinical guideline - a case-study. *Journal of the American Medical Informatics Association*, 2(4):238–242, 1995.

[59] V. L. Patel, T. Branch, D. Wang, M. Peleg, and A. Boxwala. Analysis of the process of encoding guidelines: a comparison of glif2 and glif3. *Methods Inf Med*, 41:105–13, 2002.

[60] Vimla L. Patel, Vanessa G. Allen, Jose F. Arocha, and Edward H. Shortliffe. Representing clinical guidelines in glif: individual and collaborative expertise. *Journal of the American Medical Informatics Association*, 5(5):467–483, 1998.

[61] Vimla L. Patel, José F. Arocha, Melissa Diermeier, Robert A. Greenes, and Edward H. Shortliffe. Methods of cognitive analysis to support the design and evaluation of biomedical systems: the case of clinical practice guidelines. *Journal of Biomedical Informatics*, 34(1):52–66, 2001. doi: DOI: 10.1006/jbin.2001.1002.

[62] M. Barnes and G. O. Barnett. An architecture for a distributed guideline server. *Proc Annu Symp Comput Appl Med Care*, pages 233–7, 1995.

[63] A. Bouffier and T. Poibeau. Automatically restructuring practice guidelines using the gem dtd, 2007.

[64] K. M. Hrabak, J. R. Campbell, S. W. Tu, R. McClure, and R. T. Weida. Creating interoperable guidelines: requirements of vocabulary standards in immunization decision support. *Stud Health Technol Inform*, 129(Pt 2):930–4, 2007.

[65] D. F. Lobach and N. Kerner. A systematic process for converting text-based guidelines into a linear algorithm for electronic implementation. *Proc AMIA Symp*, pages 507–11, 2000.

[66] M. Peleg, D. Wang, A. Fodor, S. Keren, and E. Karnieli. Lessons learned from adapting a generic narrative diabetic-foot guideline to an institutional decision-support system. *Stud Health Technol Inform*, 139:243–52, 2008.

[67] F. A. Sonnenberg, C. G. Hagerty, J. Acharya, D. S. Pickens, and C. A. Kulikowski. Vocabulary requirements for implementing clinical guidelines in an electronic medical record: a case study. *AMIA Annu Symp Proc*, pages 709–13, 2005.

[68] P. Dadam, M. Reichert, and K. Kuhn. Clinical workflows - the killer application for process-oriented information systems? In *Proc. 4th Int'l Conf. on Business Information Systems (BIS '00)*, pages 36–59, Poznan, Poland, 2000.

[69] S. Quaglini, M. Stefanelli, A. Cavallini, G. Micieli, C. Fassino, and C. Mossa. Guideline-based careflow systems. *Artificial Intelligence in Medicine*, 20(1):5–22, 2000. doi: DOI: 10.1016/S0933-3657(00)00050-6.

[70] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos. Emergency healthcare process automation using workflow technology and web services. *Med Inform Internet Med*, 28(3):195–207, 2003.

[71] Silvana Quaglini, Mario Stefanelli, Giordano Lanzola, Vincenzo Caporusso, and Silvia Panzarasa. Flexible guideline-based patient careflow systems. *Artificial Intelligence in Medicine*, 22(1):65–80, 2001. doi: DOI: 10.1016/S0933-3657(00)00100-7.

[72] Jiangbo Dang, Amir Hedayati, Ken Hampel, and Candemir Toklu. An ontological knowledge framework for adaptive medical workflow. *Journal of Biomedical Informatics*, 41(5):829–836, 2008. doi: DOI: 10.1016/j.jbi.2008.05.012.

[73] Katharina Kaiser and Silvia Miksch. Versioning computer-interpretable guidelines: Semi-automatic modeling of 'living guidelines' using an information extraction method. *Artificial Intelligence in Medicine*, 46(1):55–66, 2009. doi: DOI: 10.1016/j.artmed.2008.08.009.

[74] Giorgio Leonardi, Silvia Panzarasa, Silvana Quaglini, Mario Stefanelli, and Wil M. P. van der Aalst. Interacting agents through a web-based health serviceflow manage-

ment system. *Journal of Biomedical Informatics*, 40(5):486–499, 2007. doi: DOI: 10.1016/j.jbi.2006.12.002.

[75] F. Malamateniou and G. Vassilacopoulos. Developing a virtual patient record using xml and web-based workflow technologies. *International Journal of Medical Informatics*, 70(2-3):131–139, 2003. doi: DOI: 10.1016/S1386-5056(03)00039-X.

[76] Sameer Malhotra, Desmond Jordan, Edward Shortliffe, and Vimla L. Patel. Workflow modeling in critical care: Piecing together your own puzzle. *Journal of Biomedical Informatics*, 40(2):81–92, 2007. doi: DOI: 10.1016/j.jbi.2006.06.002.

[77] S. Panzarasa, S. Maddè, S. Quaglini, C. Pistarini, and M. Stefanelli. Evidence-based careflow management systems: the case of post-stroke rehabilitation. *Journal of Biomedical Informatics*, 35(2):123–139, 2002. doi: DOI: 10.1016/S1532-0464(02)00505-1.

[78] S. Panzarasa and M. Stefanelli. Workflow management systems for guideline implementation. *Neurol Sci*, 27 Suppl 3:S245–9, 2006.

[79] Federico Cabitza, Marcello Sarini, and Carla Simone. Providing awareness through situated process maps: the hospital care case. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 41–50, 2007. http://doi.acm.org/10.1145/1316624.1316631.

[80] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. van der Aalst. Process mining techniques: an application to stroke care. *Stud Health Technol Inform*, 136:573–8, 2008.

[81] G. Russello, C. Dong, N. Dulay, and Ieee. Personalising situated workflow systems for pervasive healthcare applications. In *2nd International Conference on Pervasive Computing Technologies for Healthcare*, pages 173–177. IEEE, New York, 2008.

[82] Annette ten Teije, Mar Marcos, Michel Balser, Joyce van Croonenborg, Christoph Duelli, Frank van Harmelen, Peter Lucas, Silvia Miksch, Wolfgang Reif, Kitty

Rosenbrand, and Andreas Seyfang. Improving medical protocols by formal methods. *Artificial Intelligence in Medicine*, 36(3):193–209, 2006. doi: DOI: 10.1016/j.artmed.2005.10.006.

[83] L. Allart, C. Vilhelm, H. Mehdaoui, H. Hubert, B. Sarrazin, D. Zitouni, M. Lemdani, and P. Ravaux. An architecture for online comparison and validation of processing methods and computerized guidelines in intensive care units. *Computer Methods and Programs in Biomedicine*, 93(1):93–103, 2009. doi: DOI: 10.1016/j.cmpb.2008.07.012.

[84] Rianne Bindels, Paul A. de Clercq, Ron A. G. Winkens, and Arie Hasman. A test ordering system with automated reminders for primary care based on practice guidelines. *International Journal of Medical Informatics*, 58-59:219–233, 2000. doi: DOI: 10.1016/S1386-5056(00)00089-7.

[85] Alessio Bottrighi, Laura Giordano, Gianpaolo Molino, Stefania Montani, Paolo Terenziani, and Mauro Torchio. Adopting model checking techniques for clinical guidelines verification. *Artificial Intelligence in Medicine*, 48(1):1–19, 2009. doi: DOI: 10.1016/j.artmed.2009.09.003.

[86] Isabelle Colombet, Angel-Ricardo Aguirre-Junco, Sylvain Zunino, Marie-Christine Jaulent, Laurence Leneveut, and Gilles Chatellier. Electronic implementation of guidelines in the esper system: A knowledge specification method. *International Journal of Medical Informatics*, 74(7-8):597–604, 2005. doi: DOI: 10.1016/j.ijmedinf.2005.05.001.

[87] G. Duftschmid and S. Miksch. Knowledge-based verification of clinical guidelines by detection of anomalies. *Artificial Intelligence in Medicine*, 22(1):23–41, 2001. doi: DOI: 10.1016/S0933-3657(00)00098-1.

[88] Donald W. Miller Jr, Sandra J. Frawley, and Perry L. Miller. Using semantic constraints to help verify the completeness of a computer-based clinical guideline

for childhood immunization. *Computer Methods and Programs in Biomedicine*, 58(3):267–280, 1999. doi: DOI: 10.1016/S0169-2607(98)00090-X.

[89] P. L. Miller, S. J. Frawley, and F. G. Sayward. Informatics issues in the national dissemination of a computer-based clinical guideline: A case study in childhood immunization. *Journal of the American Medical Informatics Association*, pages 580–584, 2000.

[90] P. L. Miller, S. J. Frawley, and F. G. Sayward. Maintaining and incrementally revalidating a computer-based clinical guideline: A case study. *Journal of Biomedical Informatics*, 34(2):99–111, 2001.

[91] P. L. Miller. Domain-constrained generation of clinical condition sets to help test computer-based clinical guidelines. *J Am Med Inform Assoc*, 8:131–145, 2001.

[92] J. Fox and J. Bury. A quality and safety framework for point-of-care clinical guidelines. *Proc AMIA Symp*, pages 245–9, 2000.

[93] Yuval Shahar, Silvia Miksch, and Peter Johnson. The asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. *Artificial Intelligence in Medicine*, 14(1-2):29–51, 1998. doi: DOI: 10.1016/S0933-3657(98)00015-3.

[94] R. N. Shiffman, G. Michel, A. Essaihi, and E. Thornquist. Bridging the guideline implementation gap: a systematic, document-centered approach to guideline implementation. *J Am Med Inform Assoc*, 11(5):418–26, 2004.

[95] J. Stausberg, H. Bilir, C. Waydhas, and S. Ruchholz. Guideline validation in multiple trauma care through business process modeling. *International Journal of Medical Informatics*, 70:301–307, 2003.

[96] D. Wang, M. Peleg, S. W. Tu, A. A. Boxwala, O. Ogunyemi, Q. Zeng, R. A. Greenes, V. L. Patel, and E. H. Shortliffe. Design and implementation of the glif3 guideline execution engine. *J Biomed Inform*, 37(5):305–18, 2004.

[97] Adam Wright and Dean F. Sittig. A framework and model for evaluating clinical decision support architectures. *Journal of Biomedical Informatics*, 41(6):982–990, 2008. doi: DOI: 10.1016/j.jbi.2008.03.009.

[98] Jeeyae Choi, Leanne M. Currie, Dongwen Wang, and Suzanne Bakken. Encoding a clinical practice guideline using guideline interchange format: A case study of a depression screening and management guideline. *International Journal of Medical Informatics*, 76(Supplement 2):S302–S307, 2007. doi: DOI: 10.1016/j.ijmedinf.2007.05.011.

[99] Rick Goud, Arie Hasman, and Niels Peek. Development of a guideline-based decision support system with explanation facilities for outpatient therapy. *Computer Methods and Programs in Biomedicine*, 91(2):145–153, 2008. doi: DOI: 10.1016/j.cmpb.2008.03.006.

[100] Paolo Terenziani, Gianpaolo Molino, and Mauro Torchio. A modular approach for representing and executing clinical guidelines. *Artificial Intelligence in Medicine*, 23(3):249–276, 2001. doi: DOI: 10.1016/S0933-3657(01)00087-2.

[101] Arnost Veselý, Jana Zvárová, Jan Peleska, David Buchtela, and Zdenek Anger. Medical guidelines presentation and comparing with electronic health record. *International Journal of Medical Informatics*, 75(3-4):240–245, 2006. doi: DOI: 10.1016/j.ijmedinf.2005.07.016.

[102] Paul A. de Clercq, Arie Hasman, Johannes A. Blom, and Hendrikus H. M. Korsten. Design and implementation of a framework to support the development of clinical guidelines. *International Journal of Medical Informatics*, 64(2-3):285–318, 2001. doi: DOI: 10.1016/S1386-5056(01)00189-7.

[103] G. B. Laleci and A. Dogac. A semantically enriched clinical guideline model enabling deployment in heterogeneous healthcare environments. *IEEE Trans Inf Technol Biomed*, 13(2):263–73, 2009.

[104] S. M. Maviglia, R. D. Zielstorff, M. Paterno, J. M. Teich, D. W. Bates, and G. J.

Kuperman. Automating complex guidelines for chronic disease: lessons learned. *J Am Med Inform Assoc*, 10(2):154–65, 2003.

[105] T. Burkle, T. Baur, and N. Hoss. Clinical pathways development and computer support in the epr: lessons learned. *Stud Health Technol Inform*, 124:1025–30, 2006.

[106] A. Daniyal, S. R. Abidi, and S. S. Abidi. Computerizing clinical pathways: ontology-based modeling and execution. *Stud Health Technol Inform*, 150:643–7, 2009.

[107] Kris Verlaenen, Wouter Joosen, and Pierre Verbaeten. Arriclides: an architecture integrating clinical decision support models. In *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 135c–135c, 2007.

[108] Shunji Wakamiya and Kazunobu Yamauchi. What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*, 78(8):543–550, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.03.003.

[109] S. Quaglini, M. Grandi, P. Baiardi, M. C. Mazzoleni, C. Fassino, G. Franchi, and S. Melino. A computerized guideline for pressure ulcer prevention. *Int J Med Inform*, 58-59:207–17, 2000.

[110] Kim M. Unertl, Matthew B. Weinger, Kevin B. Johnson, and Nancy M. Lorenzi. Describing and modeling workflow and information flow in chronic disease care. *Journal of the American Medical Informatics Association*, 16(6):826–836, 2009. doi: DOI: 10.1197/jamia.M3000.

[111] S. Wakamiya and K. Yamauchi. A new approach to systematization of the management of paper-based clinical pathways. *Comput Methods Programs Biomed*, 82(2):169–76, 2006.

[112] C. J. Wallace, S. Bigelow, X. Xu, and L. Elstein. Collaborative practice: usability of text-based, electronic patient care guidelines. *Comput Inform Nurs*, 25(1):39–44, 2007.

[113] V. Anand, P. G. Biondich, G. Liu, M. Rosenman, and S. M. Downs. Child health improvement through computer automation: the chica system. *Stud Health Technol Inform*, 107(Pt 1):187–91, 2004.

[114] John Fox, Alyssa Alabassi, Vivek Patkar, Tony Rose, and Elizabeth Black. An ontological approach to modelling tasks and goals. *Computers in Biology and Medicine*, 36(7-8):837–856, 2006. doi: DOI: 10.1016/j.compbiomed.2005.04.011.

[115] Georg Duftschmid, Silvia Miksch, and Walter Gall. Verification of temporal scheduling constraints in clinical practice guidelines. *Artificial Intelligence in Medicine*, 25(2):93–121, 2002.

[116] Adela Grando, Mor Peleg, and David Glasspool. A goal-oriented framework for specifying clinical guidelines and handling medical errors. *Journal of Biomedical Informatics*, 43(2):287–99, 2010. doi: DOI: 10.1016/j.jbi.2009.11.006.

[117] Y. Shahar, S. Miksch, and P. Johnson. An intention-based language for representing clinical guidelines. *Proc AMIA Annu Fall Symp*, pages 592–6, 1996.

[118] Luca Anselma, Paolo Terenziani, Stefania Montani, and Alessio Bottrighi. Towards a comprehensive treatment of repetitions, periodicity and temporal constraints in clinical guidelines. *Artificial Intelligence in Medicine*, 38(2):171–195, 2006. doi: DOI: 10.1016/j.artmed.2006.03.007.

[119] Mercedes Argüello Casteleiro and Jose Julio Des Diz. Clinical practice guidelines: a case study of combining owl-s, owl, and swrl. *Knowledge-Based Systems*, 21(3):247–255, 2008. doi: DOI: 10.1016/j.knosys.2007.11.008.

[120] A. Seyfang and S. Miksch. Advanced temporal data abstraction for guideline execution. *Stud Health Technol Inform*, 139:263–72, 2004.

[121] A. Seyfang, M. Paesold, P. Votruba, and S. Miksch. Improving the execution of clinical guidelines and temporal data abstraction high-frequency domains. *Stud Health Technol Inform*, 139:263–72, 2008.

[122] Shobha Phansalkar, Charlene R. Weir, Alan H. Morris, and Homer R. Warner. Clinicians' perceptions about use of computerized protocols: A multicenter study. *International Journal of Medical Informatics*, 77(3):184–193, 2008. doi: DOI: 10.1016/j.ijmedinf.2007.02.002.

[123] Rick Goud, Mariette van Engen-Verheul, Nicolette F. de Keizer, Roland Bal, Arie Hasman, Irene M. Hellemans, and Niels Peek. The effect of computerized decision support on barriers to guideline implementation: A qualitative study in outpatient cardiac rehabilitation. *International Journal of Medical Informatics*, 79(6):430–7, 2010. doi: DOI: 10.1016/j.ijmedinf.2010.03.001.

[124] S. Chu. Computerised clinical pathway as process quality improvement tool. In V. L. Patel, R. Rogers, and R. Haux, editors, *Medinfo 2001: Proceedings of the 10th World Congress on Medical Informatics, Pts 1 and 2*, volume 84 of *Studies in Health Technology and Informatics*, pages 1135–1139. I O S Press, Amsterdam, 2001.

[125] Mercedes Argüello Casteleiro, Julio Des, Maria Jesus Fernandez Prieto, Rogelio Perez, and Hilary Paniagua. Executing medical guidelines on the web: Towards next generation healthcare. *Knowledge-Based Systems*, 22(7):545–551, 2009. doi: DOI: 10.1016/j.knosys.2008.10.003.

[126] W. M. Tierney, J. M. Overhage, B. Y. Takesue, L. E. Harris, M. D. Murray, D. L. Vargo, and C. J. McDonald. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. *J Am Med Inform Assoc*, 2(5):316–22, 1995.

[127] Mor Peleg and Samson W. Tu. Design patterns for clinical guidelines. *Artificial Intelligence in Medicine*, 47(1):1–24, 2009. doi: DOI: 10.1016/j.artmed.2009.05.004.

[128] K. M. Unertl, L. L. Novak, K. B. Johnson, and N. M. Lorenzi. Traversing the many paths of workflow research: developing a conceptual framework of workflow terminology through a systematic literature review. *J Am Med Inform Assoc*, 17(3):265–273, 2010.

[129] Zahra Niazkhani, Habibollah Pirnejad, Marc Berg, and Jos Aarts. The impact of

computerized provider order entry systems on inpatient clinical workflow: a literature review. *Journal of the American Medical Informatics Association*, 16(4):539–549, 2009. doi: DOI: 10.1197/jamia.M2419.

[130] Gillian Hardstone, Mark Hartswood, Rob Procter, Alex Voss, and Gwyneth Rees. Supporting informality: team working and integrated care records. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 142–151, 2004.

[131] Jason J. Saleem, Alissa L. Russ, Connie F. Justice, Heather Hagg, Patricia R. Ebright, Peter A. Woodbridge, and Bradley N. Doebbeling. Exploring the persistence of paper with the electronic health record. *International Journal of Medical Informatics*, 78(9):618–628, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.04.001.

[132] P.A. Buhler and J.M. Vidal. Towards adaptive workflow enactment using multiagent systems. *Information Technology and Management*, 6(1):61–87, 2005.

[133] Christian W. Guenther, Manfred Reichert, and Wil M.P. van der Aalst. Supporting flexible processes with adaptive workflow and case handling. In *Proceedings WETICE'08, 3rd IEEE Workshop on Agile Cooperative Process-aware Information Systems (ProGility'08)*, Rome, Italy, 2008.

[134] S. R. Abidi and H. Chen. Adaptable personalized care planning via a semantic web framework. In *20th Intl Cong European Fed for Medical Informatics Maastricht*, Maastricht, 2006. IOS Press.

[135] K.F. Hurley and S. R. Abidi. Ontology engineering to model clinical pathways: Towards the computerization and execution of clinical pathways. In *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pages 536–541, Maribor, Slovenia, 2007.

[136] Hai H. Wang, Natasha Noy, Alan Rector, Mark Musen, Timothy Redmond, Daniel Rubin, Samson Tu, Tania Tudorache, Nick Drummond, Matthew Horridge, and

Julian Sedenberg. Frames and owl side by side. In *10th International Protege Conference*, Budapest, Hungary, 2007.

[137] Leila Ahmadian, Ronald Cornet, and Nicolette F. de Keizer. Facilitating pre-operative assessment guidelines representation using snomed ct. *Journal of Biomedical Informatics*, in press, 2010.

[138] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates. Grand challenges in clinical decision support. *J Biomed Inform*, 41(2):387–392, Apr 2008.

[139] Anna Hristoskova, Dieter Moeyersoon, Sofie Van Hoecke, Stijn Verstichel, Johan Decruyenaere, and Filip De Turck. Dynamic composition of medical support services in the icu: Platform and algorithm design details. *Computer Methods and Programs in Biomedicine*, 100(3):248–264, 2010.

[140] Safe and Sound. Consensus on project objectives, 2009. http://www. clinicalfutures.org.uk/consensus.

[141] Arun Sen, Amarnath Banerjee, Atish P. Sinha, and Manish Bansal. Clinical decision support: Converging toward an integrated architecture. *Journal of Biomedical Informatics*, (0):–, 2012.

References for Chapter 4

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2 edition, 2008.

[2] M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 999888:2493–2537, November 2011.

[4] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[5] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada, 1998.

[6] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[7] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[8] J. D. Patrick, D. H. Nguyen, Y. Wang, and M. Li. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc*, 18(5):574–579, 2011.

[9] Arnold W. Pratt and Milos G. Pacak. Automated processing of medical english. In *Proceedings of the 1969 conference on Computational linguistics*, COLING '69, pages 1–23, Stroudsburg, PA, USA, 1969. Association for Computational Linguistics.

[10] N. Sager, M. Lyman, N. T. Nhan, and L. J. Tick. Medical language processing: applications to patient data representation and automatic encoding. *Methods Inf Med*, 34(1-2):140–146, Mar 1995.

[11] Jon Patrick, Yefeng Wang, and Peter Budd. An automated system for conversion of clinical notes into snomed clinical terminology. In *Proceedings of the fifth Australasian symposium on ACSW frontiers - Volume 68*, ACSW '07, pages 219–226, Darlinghurst, Australia, Australia, 2007. Australian Computer Society, Inc.

[12] Katharina Kaiser and Silvia Miksch. Supporting the abstraction of clinical practice guidelines using information extraction. In *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*, NLDB'10, pages 304–311, Berlin, Heidelberg, 2010. Springer-Verlag.

[13] Radu Serban, Annette ten Teije, Frank van Harmelen, Mar Marcos, and Cristina Polo-Conde. Extraction and use of linguistic patterns for modelling medical guidelines. *Artificial Intelligence in Medicine*, 39(2):137–149, 2007. doi: DOI: 10.1016/j.artmed.2006.07.012.

[14] H. Cunningham, D Maynard, K Bontcheva, and V Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

[15] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction

ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[16] M. Romauch. Coreference resolution in clinical practice guidelines. Technical report, Diplomarbeitspräsentationen der Fakultät für Informatik, Wien, 2009.

[17] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996.

[18] J. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*, 70(4):213–220, Oct 1968.

[19] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[20] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, Feb 2012.

[21] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21, 2001.

[22] N.H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A.P. Chiang, and Musen M.A. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14, 2009.

[23] William H. DeLone and Ephraim R. McLean. The delone and mclean model of information systems success: A ten-year update. *J. Manage. Inf. Syst.*, 19(4):9–30, April 2003.

References for Chapter 5

[1] P. Gooch and A. Roudsari. A tool for enhancing metamap performance when annotating clinical guideline documents with umls concepts. In *IDAMAP Workshop at the 13th Conference on Artificial Intelligence in Medicine (AIME'11)*, 2011.

[2] David McClosky, Mihai Surdeanu, and Christopher Manning. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1635, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[3] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Proceedings of the 10th Panhellenic Conference on Informatics*, 2005.

[4] B. Settles. ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[5] Lorraine Tanabe and W. John Wilbur. Tagging Gene and Protein Names in Full Text Articles. In *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, 2002.

[6] R. McDonald and F. Pereira. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinformatics*, 6(S6), 2005.

[7] R. T. McDonald, R. S. Winters, M. Mandel, Y. Jin, P. S. White, and F. Pereira. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 20(17):3249–3251, 2004.

[8] J. Gregory Caporaso, William A. Baumgartner Jr., David A. Randolph, K. Bretonnel Cohen, , and Lawrence Hunter. MutationFinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865, 2007.

[9] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21, 2001.

[10] Aronson Alan R and Lang François-Michel. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17:229–236, 2010.

[11] Yoshinobu Kano, Makoto Miwa, K. Bretonnel Cohen, Lawrence Hunter, Sophia Ananiadou, and Jun ichi Tsujii. U-compare: A modular nlp workflow construction and evaluation system. *IBM Journal of Research and Development*, 55(3):11, 2011.

[12] Leonard D'Avolio, Dina Demner-Fushman, and Wendy W. Chapman. An introduction to clinical natural language processing. In *AMIA 2011 Annual Symposium, NLM Staff Papers and Presentations*, 2011.

[13] Steven Bird, Ewan Klein, Edward Loper, and Jason Baldridge. Multidisciplinary instruction with the natural language toolkit. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, TeachCL '08, pages 62–70, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[14] H. Cunningham, D Maynard, K Bontcheva, and V Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

[15] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348, September 2004.

[16] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural

language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[17] H. Cunningham, D. Maynard, and K. Bontcheva. University of Sheffield Department of Computer Science, Sheffield, UK, 2011.

[18] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513, 2010.

[19] A. C. Browne, G. Divita, A. R. Aronson, and A. T. McCray. UMLS language and vocabulary tools. *AMIA Annu Symp Proc*, page 798, 2003.

[20] J. Zheng, W. W. Chapman, T. A. Miller, C. Lin, R. S. Crowley, and G. K. Savova. A system for coreference resolution for the clinical narrative. *J Am Med Inform Assoc*, Jan 2012.

[21] R. S. Crowley, M. Castine, K. Mitchell, G. Chavan, T. McSherry, and M. Feldman. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *J Am Med Inform Assoc*, 17(3):253–264, 2010.

[22] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310, 2001.

[23] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform*, 42(5):839–851, Oct 2009.

[24] Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy Lazarus, and Ross. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6:30, 2006.

[25] S. Meystre and P. J. Haug. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *J Biomed Inform*, 39(6):589–599, Dec 2006.

[26] M Krauthammer and G Nenadić. Term identification in the biomedical literature. *J Biomed Inf*, 37:512–526, 2004.

[27] Vijayaraghavan Bashyam and Ricky K. Taira. Identifying anatomical phrases in clinical reports by shallow semantic parsing methods. In *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 210–214. IEEE, 2007.

[28] M. Davis and M. Durst. Unicode normalization forms, 2003.

[29] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[30] Katharina Kaiser and Silvia Miksch. Supporting the abstraction of clinical practice guidelines using information extraction. In *Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems*, NLDB'10, pages 304–311, Berlin, Heidelberg, 2010. Springer-Verlag.

[31] Katharina Kaiser, Andreas Seyfang, and Silvia Miksch. Identifying actions described in clinical practice guidelines using semantic relations. In *Proc. of the KR4HC 2010 - 2nd International Workshop on Knowledge Representation for Health-Care in conjunction with the European Conference on Artificial Intelligence (ECAI 2010)*, August 17th 2010.

[32] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. Timeml: Robust

specification of event and temporal expressions in text. In *New Directions in Question Answering*, pages 28–34. AAAI Press, 2003.

[33] S. Agarwal and H. Yu. Detecting hedge cues and their scope in biomedical text with conditional random fields. *J Biomed Inform*, 43(6):953–961, Dec 2010.

[34] Ö Uzuner, Y Juo, and P Szolovits. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14(5):550–63, 2007.

[35] Rob Koeling, A. Rosemary Tate, and John A. Carroll. Automatically estimating the incidence of symptoms recorded in gp free text notes. In *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, MIXHS '11, pages 43–50, New York, NY, USA, 2011. ACM.

[36] Angel X. Chang and Christopher Manning. Sutime: A library for recognizing and normalizing time expressions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[37] J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656, Lancaster, March 2003.

[38] Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753. ELRA, 2012.

[39] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772, 2009. doi: DOI: 10.1016/j.jbi.2009.08.007.

[40] O. Young, Y. Shahar, Y. Liel, E. Lunenfeld, G. Bar, E. Shalom, S. B. Martins, L. T. Vaszar, T. Marom, and M. K. Goldstein. Runtime application of hybrid-asbru clinical guidelines. *J Biomed Inform*, 40(5):507–26, 2007.

References for Chapter 6

[1] P. Gooch. Systematic identification and correction of spelling errors in the foundational model of anatomy. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, SWAT4LS '11, pages 34–35, New York, NY, USA, 2012. ACM.

[2] C. Rosse and J.V.L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36:478–500, 2003.

[3] Aronson Alan R and Lang François-Michel. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17:229–236, 2010.

[4] A. C. Browne, G. Divita, A. R. Aronson, and A. T. McCray. UMLS language and vocabulary tools. *AMIA Annu Symp Proc*, page 798, 2003.

[5] H. M. Müller, E. E. Kenny, and P. W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11), November 2004.

[6] Alexa T. McCray, Allen C. Browne, and Dorothy L. Moore. The semantic structure of neo-classical compounds. Technical report, National Libnury of Medicine, 1988.

[7] P. Gooch and A. Roudsari. Automated recognition and post-coordination of complex clinical terms. *Stud Health Technol Inform*, 164:8–12, 2011.

[8] Roberto Navigli, Paola Velardi, and Stefano Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second inter-*

national joint conference on Artificial Intelligence - Volume Volume Three, IJCAI'11, pages 1872–1877. AAAI Press, 2011.

[9] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, 2004.

[10] M.A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Conference on Computational Linguistics (COLING'92)*, Nantes, France, 1992. Association for Computational Linguistics.

[11] Diana Maynard, Adam Funk, and Wim Peters. Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proceedings of WOP2009 collocated with ISWC2009*, volume 516. CEUR-WS.org, November 2009.

[12] M. Dai, N.H. Shah, W. Xuan, M.A. Musen, S.J. Watson, B. Athey, and F. Meng. An efficient solution for mapping free text to ontology terms. In *AMIA Summit on Translational Bioinformatics*, 2008.

[13] Wojciech Szpankowski. Patricia tries again revisited. *J. ACM*, 37(4):691–711, October 1990.

[14] N.H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A.P. Chiang, and Musen M.A. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14, 2009.

[15] E. Jonquet. Open biomedical annotator and mgrep source code.

[16] Chiara Del Vescovo, Damian D. G. Gessler, Pavel Klinov, Bijan Parsia, Ulrike Sattler, Thomas Schneider, and Andrew Winget. Decomposition and modular structure of bioportal ontologies. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 130–145, Berlin, Heidelberg, 2011. Springer-Verlag.

[17] Tuanjie Tong, Yugyung Lee, and Deendayal Dinakarpandian. Go-words: An entropic approach to semantic decomposition of gene ontology terms. 2008.

[18] Sophia Ananiadou. A methodology for automatic term recognition. In *Proceedings of the 15th conference on Computational linguistics*, volume 2, pages 1034–1038, Kyoto, Japan, 1994. Association for Computational Linguistics.

[19] J. D. Osborne, J. Flatow, M. Holko, S. M. Lin, W. A. Kibbe, L. J. Zhu, M. I. Danila, G. Feng, and R. L. Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10 Suppl 1:S6, 2009.

[20] H. Cunningham, D. Maynard, and K. Bontcheva. University of Sheffield Department of Computer Science, Sheffield, UK, 2011.

[21] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–465, 2011.

[22] A. T. McCray, A. Burgun, and O. Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. *Stud Health Technol Inform*, 84(Pt 1):216–220, 2001.

[23] S. Pyysalo, T. Ohta, and S. Ananiadou. Anatomical entity recognition with open biomedical ontologies. In *Proceedings of the Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*, 2011.

[24] V. Bashyam and R. K. Taira. Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA. *AMIA Annu Symp Proc*, pages 26–30, 2005.

References for Chapter 7

[1] P. Gooch. Badrex: In situ expansion and coreference of biomedical abbreviations using dynamic regular expressions. Technical report, City University London, 2012.

[2] M. Torii, Z. Z. Hu, M. Song, C. H. Wu, and H. Liu. A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics*, 8 Suppl 9:S5, 2007.

[3] S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658–3664, Sep 2005.

[4] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–462, 2003.

[5] H. Ao and T. Takagi. ALICE: an algorithm to extract abbreviations from MEDLINE. *J Am Med Inform Assoc*, 12(5):576–586, 2005.

[6] Mark Stevenson, Yikun Guo, Abdulaziz Al Amri, and Robert Gaizauskas. Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 71–79, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[7] M. Stevenson, E. Agirre, and A. Soroa. Exploiting domain information for Word Sense Disambiguation of medical documents. *J Am Med Inform Assoc*, 19(2):235–240, 2012.

[8] H. Cunningham, D Maynard, K Bontcheva, and V Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

[9] J et al. Pustejovsky. Medstract: natural language tools for mining the biobibliome, 2002. [Accessed 10 May 2012].

[10] Y. Xu, Z. Wang, Y. Lei, Y. Zhao, and Y. Xue. MBA: a literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics*, 10:14, 2009.

References for Chapter 8

[1] P. Gooch and A. Roudsari. Lexical patterns, features and knowledge resources for coreference resolution in clinical notes. *Journal of Biomedical Informatics*, 2012. doi: DOI: 10.1016/j.jbi.2012.02.012.

[2] C. Gasperin. Statistical anaphora resolution in biomedical texts. Technical report, University of Cambridge Computer Laboratory Technical Report No. 764, 2009. http:// www.cl.cam.ac.uk/techreports/UCAM-CL-TR-764.pdf.

[3] van Deemter K. and Kibble R. On coreferring: coreference in muc and related annotation schemes. *Comput Linguist*, 26:629–37, 2001.

[4] A. Rahman and V. Ng. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution. *J Artif Intell Res*, 40:469–521, 2011.

[5] V. Ng. Supervised noun phrase coreference research: the first fifteen years. In *Proceedings of the 48th annual meeting of the association for computational linguistics 2010*, page 1396–1411, Stroudsburg, PA, 2010. Association for Computational Linguistics.

[6] Lee H., Peirsman Y., Chang A., Chambers N., Surdeanu M., and Jurafsky D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the CoNLL-2011 shared task, 2011*, 2011.

[7] N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 602–610, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[8] J. Zheng, W.W. Chapman, R.S. Crowley, and G.K. Savova. Coreference resolution: a review of general methodologies and applications in the clinical domain. *J Biomed Inform*, 44(6):1113–1122, Dec 2011.

[9] Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, and J. Smith. Bart: a modular toolkit for coreference resolution. In *Proceedings of the 6th international conference on language resources and evaluation (LREC 2008)*, 2008.

[10] D. Hinote, D. Ramirez, and Ping Chen. A comparative study of co-reference resolution in clinical text. In *The fifth i2b2/VA/cincinnati workshop on challenges in natural language processing for clinical data*, Washington, DC, October, 2011.

[11] M. Romauch. Coreference resolution in clinical practice guidelines. Technical report, Diplomarbeitspräsentationen der Fakultät für Informatik, Wien, 2009.

[12] T-Y. He. Coreference resolution on entities and events for hospital discharge summaries. thesis (m. eng.). Technical report, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2007.

[13] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, Feb 2012.

[14] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–465, 2011.

[15] O. Uzuner. 2011 i2b2/va coreference annotation guidelines for the clinical domain. [Accessed 31.01.12].

[16] A. Bodnari. Coreference resolution evaluation script, 1.6.3, June 2011. https://www.i2b2.org/NLP/Coreference/assets/coreference_evaluation_metrics.zip.

[17] A. Bagga and B. Baldwin. Algorithms for scoring coreference chains. In *Proceedings of the 1st international conference on language resources and evaluation*, pages 563—66, Granada, Spain, 1998.

[18] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 45–52, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[19] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 25–32, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[20] M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510, 2011.

[21] J. Zheng, W. W. Chapman, T. A. Miller, C. Lin, R. S. Crowley, and G. K. Savova. A system for coreference resolution for the clinical narrative. *J Am Med Inform Assoc*, Jan 2012.

[22] J. Cai and M. Strube. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 28–36, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[24] Hoifung Poon and Pedro Domingos. Joint unsupervised coreference resolution with markov logic. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 650–659, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[25] H. Cunningham, D. Maynard, and K. Bontcheva. University of Sheffield Department of Computer Science, Sheffield, UK, 2011.

[26] H. Cunningham, D Maynard, K Bontcheva, and V Tablan. Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 2002.

[27] N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, page 789–97, Columbus, OH, 2008. Association for Computational Linguistics.

[28] C. Rosse and J.V.L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36:478–500, 2003.

[29] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, pages 17–21, 2001.

[30] P. Gooch and A. Roudsari. A tool for enhancing metamap performance when annotating clinical guideline documents with umls concepts. In *IDAMAP Workshop at the 13th Conference on Artificial Intelligence in Medicine (AIME'11)*, 2011.

[31] Lexical Systems Group. Gspell. [Accessed 31.01.12].

[32] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[33] W. Cohen, P. Ravikumar, S. Fienberg, and K. Rivard. Secondstring: an open-source java-based package of approximate string-matching techniques. [accessed 31.01.12].

[34] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration*, pages 73–78, August 2003.

[35] William E. Winkler. String comparator metrics and enhanced decision rules in the

fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359, 1990.

[36] Alvaro E. Monge and Charles Elkan. The field matching problem: Algorithms and applications. In *Proceedings of the second international conference on knowledge discovery and data mining*, pages 267–270, 1996.

[37] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

References for Chapter 9

[1] R. Parasuraman and D. H. Manzey. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*, 52(3):381–410, Jun 2010.

[2] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, and B. R. South. Evaluating the state of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, Feb 2012.

[3] B. Rink, K. Roberts, and S. M. Harabagiu. A supervised framework for resolving coreference in clinical records. *J Am Med Inform Assoc*, 19(5):875–882, Sep 2012.

[4] David Isern and Antonio Moreno. Computer-based execution of clinical guidelines: A review. *International Journal of Medical Informatics*, 77(12):787–808, 2008. doi: DOI: 10.1016/j.ijmedinf.2008.05.010.

[5] Nataliya Mulyar, Wil M. P. van der Aalst, and Mor Peleg. A pattern-based analysis of clinical computer-interpretable guideline modeling languages. *Journal of the American Medical Informatics Association*, 14(6):781–787, 2007. doi: DOI: 10.1197/jamia.M2389.

[6] Robert A. Greenes. Features of computer-based clinical decision support. In R. A. Greenes, editor, *Clinical Decision Support: The Road Ahead*, pages 79–107. Academic Press, Burlington, 2007. doi: DOI: 10.1016/B978-012369377-8/50004-0.

[7] David R. Sutton and John Fox. The syntax and semantics of the proforma guideline modeling language. *Journal of the American Medical Informatics Association*, 10(5):433–443, 2003. doi: DOI: 10.1197/jamia.M1264.

[8] William H. DeLone and Ephraim R. McLean. The delone and mclean model of information systems success: A ten-year update. *J. Manage. Inf. Syst.*, 19(4):9–30, April 2003.

[9] P. Gooch. Systematic identification and correction of spelling errors in the foundational model of anatomy. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, SWAT4LS '11, pages 34–35, New York, NY, USA, 2012. ACM.

[10] C. Manning. Overview: Statistics focused approaches to NLP. In *Natural Language Processing: State of the Art, Future Directions and Applications for Enhancing Clinical Decision-Making, April 23–24*, 2012.

[11] Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[12] Angel X. Chang and Christopher Manning. Sutime: A library for recognizing and normalizing time expressions. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[13] Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753. ELRA, 2012.

[14] B. Kaplan. Evaluating informatics applications–some alternative approaches: the-

ory, social interactionism, and call for methodological pluralism. *Int J Med Inform*, 64(1):39–56, Nov 2001.

[15] S. Chu and B. Cesnik. Improving clinical pathway design: lessons learned from a computerised prototype. *International Journal of Medical Informatics*, 51(1):1–11, 1998.

[16] M. Stein. The map of medicine® - an innovative knowledge management tool. In *AMIA 2006 Symposium Proceedings*, page 1196, 2006.

[17] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates. Grand challenges in clinical decision support. *J Biomed Inform*, 41(2):387–392, Apr 2008.

[18] Richard N. Shiffman, George Michel, Michael Krauthammer, Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Writing clinical practice guidelines in controlled natural language. In *Proceedings of the 2009 conference on Controlled natural language*, CNL'09, pages 265–280, Berlin, Heidelberg, 2010. Springer-Verlag.

[19] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[20] S. Agarwal and H. Yu. Detecting hedge cues and their scope in biomedical text with conditional random fields. *J Biomed Inform*, 43(6):953–961, Dec 2010.

[21] K. B. Wagholikar, K. L. Maclaughlin, M. R. Henry, R. A. Greenes, R. A. Hankey, H. Liu, and R. Chaudhry. Clinical decision support with automated text processing for cervical cancer screening. *J Am Med Inform Assoc*, 19(5):833–839, Sep 2012.

[22] Broadbent M. Using GATE to extract information from clinical records for research purposes. Technical report, Kings College London, 2011.

[23] W. M. P. van der Aalst and A. J. M. M. Weijters. Process mining: a research agenda. *Comput. Ind.*, 53(3):231–244, April 2004.

[24] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W. van der Aalst. Process mining techniques: an application to stroke care. *Stud Health Technol Inform*, 136:573–8, 2008.

[25] Z. Huang, X. Lu, and H. Duan. On mining clinical pathway patterns from medical behaviors. *Artif Intell Med*, 56(1):35–50, Sep 2012.

[26] M. Lang, T. Burkle, S. Laumann, and H. U. Prokosch. Process mining for clinical workflows: challenges and current limitations. *Stud Health Technol Inform*, 136:229–234, 2008.

References for Appendix B

[1] W. M. Tierney, J. M. Overhage, B. Y. Takesue, L. E. Harris, M. D. Murray, D. L. Vargo, and C. J. McDonald. Computerizing guidelines to improve care and patient outcomes: the example of heart failure. *J Am Med Inform Assoc*, 2(5):316–22, 1995.

[2] M. Barnes and G. O. Barnett. An architecture for a distributed guideline server. *Proc Annu Symp Comput Appl Med Care*, pages 233–7, 1995.

[3] John Fox, Nicky Johns, Colin Lyons, Ali Rahmanzadeh, Richard Thomson, and Peter Wilson. Proforma: a general technology for clinical decision support systems. *Computer Methods and Programs in Biomedicine*, 54(1-2):59–67, 1997. doi: DOI: 10.1016/S0169-2607(97)00034-5.

[4] S. Chu and B. Cesnik. Improving clinical pathway design: lessons learned from a computerised prototype. *International Journal of Medical Informatics*, 51(1):1–11, 1998.

[5] S. B. Henry, K. Douglas, G. Galzagorry, A. Lahey, and W. L. Holzemer. A template-based approach to support utilization of clinical practice guidelines within an electronic health record. *J Am Med Inform Assoc*, 5(3):237–44, 1998.

[6] Donald W. Miller Jr, Sandra J. Frawley, and Perry L. Miller. Using semantic constraints to help verify the completeness of a computer-based clinical guideline for childhood immunization. *Computer Methods and Programs in Biomedicine*, 58(3):267–280, 1999. doi: DOI: 10.1016/S0169-2607(98)00090-X.

[7] Rianne Bindels, Paul A. de Clercq, Ron A. G. Winkens, and Arie Hasman. A test ordering system with automated reminders for primary care based on practice guide-

lines. *International Journal of Medical Informatics*, 58-59:219–233, 2000. doi: DOI: 10.1016/S1386-5056(00)00089-7.

[8] P. L. Miller, S. J. Frawley, and F. G. Sayward. Informatics issues in the national dissemination of a computer-based clinical guideline: A case study in childhood immunization. *Journal of the American Medical Informatics Association*, pages 580–584, 2000.

[9] S. Quaglini, M. Stefanelli, A. Cavallini, G. Micieli, C. Fassino, and C. Mossa. Guideline-based careflow systems. *Artificial Intelligence in Medicine*, 20(1):5–22, 2000. doi: DOI: 10.1016/S0933-3657(00)00050-6.

[10] S. Quaglini, M. Grandi, P. Baiardi, M. C. Mazzoleni, C. Fassino, G. Franchi, and S. Melino. A computerized guideline for pressure ulcer prevention. *Int J Med Inform*, 58-59:207–17, 2000.

[11] P. Dadam, M. Reichert, and K. Kuhn. Clinical workflows - the killer application for process-oriented information systems? In *Proc. 4th Int'l Conf. on Business Information Systems (BIS '00)*, pages 36–59, Poznan, Poland, 2000.

[12] S. Chu. Computerised clinical pathway as process quality improvement tool. In V. L. Patel, R. Rogers, and R. Haux, editors, *Medinfo 2001: Proceedings of the 10th World Congress on Medical Informatics, Pts 1 and 2*, volume 84 of *Studies in Health Technology and Informatics*, pages 1135–1139. I O S Press, Amsterdam, 2001.

[13] Paul A. de Clercq, Arie Hasman, Johannes A. Blom, and Hendrikus H. M. Korsten. Design and implementation of a framework to support the development of clinical guidelines. *International Journal of Medical Informatics*, 64(2-3):285–318, 2001. doi: DOI: 10.1016/S1386-5056(01)00189-7.

[14] Vladislav J. Mikulich, Yi-Ching A. Liu, Jennifer Steinfeldt, and David L. Schriger. Implementation of clinical guidelines through an electronic medical record: physician usage, satisfaction and assessment. *International Journal of Medical Informatics*, 63(3):169–178, 2001. doi: DOI: 10.1016/S1386-5056(01)00177-0.

[15] P. L. Miller. Domain-constrained generation of clinical condition sets to help test computer-based clinical guidelines. *J Am Med Inform Assoc*, 8:131–145, 2001.

[16] Paolo Terenziani, Gianpaolo Molino, and Mauro Torchio. A modular approach for representing and executing clinical guidelines. *Artificial Intelligence in Medicine*, 23(3):249–276, 2001. doi: DOI: 10.1016/S0933-3657(01)00087-2.

[17] S. Panzarasa, S. Maddè, S. Quaglini, C. Pistarini, and M. Stefanelli. Evidence-based careflow management systems: the case of post-stroke rehabilitation. *Journal of Biomedical Informatics*, 35(2):123–139, 2002. doi: DOI: 10.1016/S1532-0464(02)00505-1.

[18] F. Malamateniou and G. Vassilacopoulos. Developing a virtual patient record using xml and web-based workflow technologies. *International Journal of Medical Informatics*, 70(2-3):131–139, 2003. doi: DOI: 10.1016/S1386-5056(03)00039-X.

[19] S. M. Maviglia, R. D. Zielstorff, M. Paterno, J. M. Teich, D. W. Bates, and G. J. Kuperman. Automating complex guidelines for chronic disease: lessons learned. *J Am Med Inform Assoc*, 10(2):154–65, 2003.

[20] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos. Emergency healthcare process automation using workflow technology and web services. *Med Inform Internet Med*, 28(3):195–207, 2003.

[21] S. A. Barretto, J. Warren, A. Goodchild, L. Bird, S. Heard, and M. Stumptner. Linking guidelines to electronic health record design for improved chronic disease management. *AMIA Annu Symp Proc*, pages 66–70, 2003.

[22] Y. Shahar, O. Young, E. Shalom, M. Galperin, A. Mayaffit, R. Moskovitch, and A. Hessing. A framework for a distributed, hybrid, multiple-ontology clinical-guideline library, and automated guideline-support tools. *J Biomed Inform*, 37(5):325–44, 2004.

[23] D. Wang, M. Peleg, S. W. Tu, A. A. Boxwala, O. Ogunyemi, Q. Zeng, R. A. Greenes, V. L. Patel, and E. H. Shortliffe. Design and implementation of the glif3 guideline execution engine. *J Biomed Inform*, 37(5):305–18, 2004.

[24] V. Anand, P. G. Biondich, G. Liu, M. Rosenman, and S. M. Downs. Child health improvement through computer automation: the chica system. *Stud Health Technol Inform*, 107(Pt 1):187–91, 2004.

[25] P. Ciccarese, E. Caffi, L. Boiocchi, S. Quaglini, and M. Stefanelli. A guideline management system. *Stud Health Technol Inform*, 107(Pt 1):28–32, 2004.

[26] P. Ciccarese, E. Caffi, S. Quaglini, and M. Stefanelli. Architectures and tools for innovative health information systems: the guide project. *Int J Med Inform*, 74(7-8):553–62, 2005.

[27] Isabelle Colombet, Angel-Ricardo Aguirre-Junco, Sylvain Zunino, Marie-Christine Jaulent, Laurence Leneveut, and Gilles Chatellier. Electronic implementation of guidelines in the esper system: A knowledge specification method. *International Journal of Medical Informatics*, 74(7-8):597–604, 2005. doi: DOI: 10.1016/j.ijmedinf.2005.05.001.

[28] W. Michalowski, R. Slowinski, S. Wilk, K. J. Farion, J. Pike, and S. Rubin. Design and development of a mobile system for supporting emergency triage. *Methods Inf Med*, 44(1):14–24, 2005.

[29] B. Seroussi, J. Bouaud, and G. Chatellier. Guideline-based modeling of therapeutic strategies in the special case of chronic diseases. *International Journal of Medical Informatics*, 74(2-4):89–99, 2005. doi: DOI: 10.1016/j.ijmedinf.2004.06.004.

[30] Wolfgang Aigner and Silvia Miksch. Carevis: Integrated visualization of computerized protocols and temporal patient data. *Artificial Intelligence in Medicine*, 37(3):203–218, 2006. doi: DOI: 10.1016/j.artmed.2006.04.002.

[31] T. Burkle, T. Baur, and N. Hoss. Clinical pathways development and computer support in the epr: lessons learned. *Stud Health Technol Inform*, 124:1025–30, 2006.

[32] L. Hayward-Rowse and T. Whittle. A pilot project to design, implement and evaluate an electronic integrated care pathway. *J Nurs Manag*, 14(7):564–71, 2006.

[33] S. Panzarasa and M. Stefanelli. Workflow management systems for guideline implementation. *Neurol Sci*, 27 Suppl 3:S245–9, 2006.

[34] Arnost Veselý, Jana Zvárová, Jan Peleska, David Buchtela, and Zdenek Anger. Medical guidelines presentation and comparing with electronic health record. *International Journal of Medical Informatics*, 75(3-4):240–245, 2006. doi: DOI: 10.1016/j.ijmedinf.2005.07.016.

[35] S. Wakamiya and K. Yamauchi. A new approach to systematization of the management of paper-based clinical pathways. *Comput Methods Programs Biomed*, 82(2):169–76, 2006.

[36] Katharina Kaiser, Cem Akkaya, and Silvia Miksch. How can information extraction ease formalizing treatment processes in clinical practice guidelines?: A method and its evaluation. *Artificial Intelligence in Medicine*, 39(2):151–163, 2007. doi: DOI: 10.1016/j.artmed.2006.07.011.

[37] R. Lenz, R. Blaser, M. Beyer, O. Heger, C. Biber, M. Bäumlein, and M. Schnabel. It support for clinical pathways: lessons learned. *International Journal of Medical Informatics*, 76(Supplement 3):S397–S402, 2007. doi: DOI: 10.1016/j.ijmedinf.2007.04.012.

[38] Giorgio Leonardi, Silvia Panzarasa, Silvana Quaglini, Mario Stefanelli, and Wil M. P. van der Aalst. Interacting agents through a web-based health serviceflow management system. *Journal of Biomedical Informatics*, 40(5):486–499, 2007. doi: DOI: 10.1016/j.jbi.2006.12.002.

[39] O. Young, Y. Shahar, Y. Liel, E. Lunenfeld, G. Bar, E. Shalom, S. B. Martins, L. T. Vaszar, T. Marom, and M. K. Goldstein. Runtime application of hybrid-asbru clinical guidelines. *J Biomed Inform*, 40(5):507–26, 2007.

[40] Sartipi, Kamran, H. Yarmand Mohammad, and G. Down Douglas. Mined-knowledge and decision support services in electronic health. In *Proceedings of the International Workshop on Systems Development in SOA Environments*, pages 10–10, 2007. http://dx.doi.org/10.1109/SDSOA.2007.9.

[41] Kris Verlaenen, Wouter Joosen, and Pierre Verbaeten. Arriclides: an architecture integrating clinical decision support models. In *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 135c–135c, 2007.

[42] D. Alexandrou, F. Xenikoudakis, and G. Mentzas. Adaptive clinical pathways with semantic web rules. In *Proceedings of the First International Conference on Health Informatics, HEALTHINF 2008*, volume 2, pages 140–147, Funchal, Madeira, Portugal, 2008. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

[43] Mercedes Argüello Casteleiro and Jose Julio Des Diz. Clinical practice guidelines: a case study of combining owl-s, owl, and swrl. *Knowledge-Based Systems*, 21(3):247–255, 2008. doi: DOI: 10.1016/j.knosys.2007.11.008.

[44] Jiangbo Dang, Amir Hedayati, Ken Hampel, and Candemir Toklu. An ontological knowledge framework for adaptive medical workflow. *Journal of Biomedical Informatics*, 41(5):829–836, 2008. doi: DOI: 10.1016/j.jbi.2008.05.012.

[45] Rick Goud, Arie Hasman, and Niels Peek. Development of a guideline-based decision support system with explanation facilities for outpatient therapy. *Computer Methods and Programs in Biomedicine*, 91(2):145–153, 2008. doi: DOI: 10.1016/j.cmpb.2008.03.006.

[46] V. Patkar and J. Fox. Clinical guidelines and care pathways: a case study applying proforma decision support technology to the breast cancer care pathway. *Stud Health Technol Inform*, 139:233–42, 2008.

[47] A. Seyfang, M. Paesold, P. Votruba, and S. Miksch. Improving the execution of clinical guidelines and temporal data abstraction high-frequency domains. *Stud Health Technol Inform*, 139:263–72, 2008.

[48] L. Allart, C. Vilhelm, H. Mehdaoui, H. Hubert, B. Sarrazin, D. Zitouni, M. Lemdani, and P. Ravaux. An architecture for online comparison and validation of processing

methods and computerized guidelines in intensive care units. *Computer Methods and Programs in Biomedicine*, 93(1):93–103, 2009. doi: DOI: 10.1016/j.cmpb.2008.07.012.

[49] Mercedes Argüello Casteleiro, Julio Des, Maria Jesus Fernandez Prieto, Rogelio Perez, and Hilary Paniagua. Executing medical guidelines on the web: Towards next generation healthcare. *Knowledge-Based Systems*, 22(7):545–551, 2009. doi: DOI: 10.1016/j.knosys.2008.10.003.

[50] G. B. Laleci and A. Dogac. A semantically enriched clinical guideline model enabling deployment in heterogeneous healthcare environments. *IEEE Trans Inf Technol Biomed*, 13(2):263–73, 2009.

[51] Mor Peleg, Aviv Shachak, Dongwen Wang, and Eddy Karnieli. Using multi-perspective methodologies to study users' interactions with the prototype front end of a guideline-based decision support system for diabetic foot care. *International Journal of Medical Informatics*, 78(7):482–493, 2009. doi: DOI: 10.1016/j.ijmedinf.2009.02.008.

[52] A. Daniyal, S. R. Abidi, and S. S. Abidi. Computerizing clinical pathways: ontology-based modeling and execution. *Stud Health Technol Inform*, 150:643–7, 2009.

[53] C. Eccher, A. Seyfang, A. Ferro, and S. Miksch. Embedding oncologic protocols into the provision of care: the oncocure project. *Stud Health Technol Inform*, 150:663–7, 2009.

[54] M. Tschopp, M. Despond, D. Grauser, J. C. Staub, and C. Lovis. Computer-based physician order entry: implementation of clinical pathways. *Stud Health Technol Inform*, 150:673–7, 2009.