LINKING ELECTRONIC HEALTH RECORDS WITH THE BIOMEDICAL LITERATURE

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2016

By
Xiao Fu
School of Computer Science
Contents

List of Tables ......................................................................................................................... 5
List of Figures .......................................................................................................................... 7
Abstract ................................................................................................................................. 10
Declaration .............................................................................................................................. 11
Copyright ............................................................................................................................... 12
Acknowledgements ............................................................................................................... 14
Abbreviations ......................................................................................................................... 15

1 Introduction ......................................................................................................................... 19
  1.1 Motivation ................................................................................................................... 19
  1.2 Research hypotheses and objectives ........................................................................... 20
  1.3 Overview of contributions ........................................................................................... 22

2 Phenotype extraction .......................................................................................................... 24
  2.1 Literature review ......................................................................................................... 24
    2.1.1 Pre-processing ..................................................................................................... 25
    2.1.2 Post-processing .................................................................................................. 27
    2.1.3 Phenotype extraction in the clinical domain ....................................................... 28
    2.1.4 Phenotype extraction in the biomedical domain ............................................... 29
4.1.2 Domain adaptation ................................................................. 103

4.2 Methodology ............................................................................ 105

4.2.1 Compositional synonym generation ...................................... 106

4.2.2 Normalisation of syntactic structures .................................... 115

4.2.3 Normalisation of acronyms and abbreviations ..................... 123

4.2.4 Normalisation of plural and singular .................................... 124

4.2.5 Individual and hybrid methods ............................................ 126

4.3 Results and discussion ........................................................... 126

4.3.1 Variants generation ............................................................. 126

4.3.2 Evaluation ............................................................................ 128

4.4 Discussion ................................................................................ 137

4.5 Summary ................................................................................ 144

5 Conclusion .................................................................................. 145

5.1 Evaluation of research objectives ............................................ 145

5.2 Future work ............................................................................ 150

References ...................................................................................... 152

Appendix A List of COPD phenotypes used to retrieve articles from the PubMed OpenAccess subset ................................................................. 172

Appendix B Finer-grained COPD phenotypes .................................. 189

B.1 Problem .................................................................................. 189

B.2 Treatment .............................................................................. 191

B.3 Test ....................................................................................... 191

Word Count: 38,210
List of Tables

Table 2.1: Examples from a pair of report and annotation files. ............................... 33
Table 2.2: Transformed annotations of “Trauma series demonstrated no evidence of a pelvic fracture.”, ........................................................................................................... 37
Table 2.3: Proximity relationships used to calculate DS. ......................................... 44
Table 2.4: Definition of TP, FP, and FN. ................................................................. 47
Table 2.5: Exact and relaxed matching results for concept extraction. ................. 51
Table 2.6: Exact and relaxed matching results for each concept type ((a) Problem; (b) Treatment; (c) Test) .......................................................................................................................... 53
Table 2.7: Number of unrecognised abbreviations. ............................................. 57
Table 3.1: Number of MIMIC clinical records of each disease and its relationship with COPD .......................................................... 65
Table 3.2: Distribution of 1000 MIMIC II clinical records................................. 69
Table 3.3: Description of collected clinical records of COPD patients............... 71
Table 3.4: PMC journals relevant to COPD. ........................................................ 72
Table 3.5: 30 top-ranked biomedical articles. ..................................................... 74
Table 3.6: Annotation scheme for capturing COPD phenotypes (*: Semantic types adapted from the PhenoCHF scheme) ................................................................. 77
Table 3.7: Examples of phenotypes represented using our annotation scheme. .... 78
Table 3.8: Argo components used in our annotation workflow and their descriptions. ............................................................................................................................ 80
Table 3.9: Numbers of occurrences of automatically generated annotations (Values in brackets indicate counts of unique instances) ................................................................. 87
Table 3.10: Number of unique concepts for each type .............................................. 89
Table 3.11: Evaluation of annotations automatically generated by the text mining-assisted workflow against gold standard data ................................................................. 91
Table 3.12: Results of 10-fold cross validation of concept recognisers, using exact matching (Performance is compared with that of the components utilised in the text mining-assisted workflow) ........................................................................................................ 93
Table 3.13: Results of evaluation using a fixed split over 1,299 paragraphs (Training set: 75% or 973 paragraphs; held-out set: 25% or 325 paragraphs), using exact matching ................................................................................................................................. 94
Table 4.1: Examples of phenotypic concepts expressed in Greek and English. ..... 110
Table 4.2: Examples of Greek elements and their expression in English. .............. 111
Table 4.3: Examples of transformation between English and Greek concepts ((a) English to Greek; (b) Greek to English) ........................................................................ 113
Table 4.4: Examples of different syntactic constructions of concepts identified by Enju................................................................................................................................. 116
Table 4.5: Output of the Enju parser for cholesterol elevation, elevation of cholesterol, elevated cholesterol, elevating cholesterol .................................................. 117
Table 4.6: Examples of syntactic transformation of some special cases .............. 123
Table 4.7: Base forms of thrombi, fasciculations, and high blood sugars ............ 125
Table 4.8: Number of unique variants generated ..................................................... 127
Table 4.9: Normalisation performance of individual methods in both domains .... 129
Table 4.10: Normalisation performance of hybrid methods in both domains ((a) Clinical domain; (b) Biomedical domain (in the ascending order of F-scores)) .... 132
Table 4.11: Normalisation performance of all our methods in terms of accuracy .. 136
Table 4.12: Comparison of our hybrid method against other approaches applied to the biomedical corpus .................................................................................................. 142
Table 4.13: Comparison of our hybrid method against other approaches applied to the clinical corpus ........................................................................................................ 143
List of Figures

Figure 1.1: Framework of linking clinical records with biomedical literature........ 23
Figure 2.1: Framework of baseline using Argo......................................................... 34
Figure 2.2: Example snippets of (a) Annotation files from the i2b2/VA challenge; (b) Converted named entity label (the first column) and byte positions (the second and the third columns) for training the NERsuite model ........................................ 36
Figure 2.3: Example snippet of test files annotated by NERsuite......................... 38
Figure 2.4: Framework of truecasing using modules in Argo................................. 39
Figure 2.5: Architecture of Argo............................................................................. 39
Figure 2.6: A screenshot of the workflow diagramming interface of Argo (Example of the GENIA Tagger component) ................................................................. 40
Figure 2.7: Truecase Asciifier module in Argo....................................................... 41
Figure 2.8: Framework of abbreviation disambiguation......................................... 43
Figure 2.9: Framework of using distributional similarity. ........................................ 45
Figure 2.10: Example of how the distributional thesaurus works. .......................... 46
Figure 2.11: Framework of the sequential hybrid annotation system..................... 48
Figure 2.12: Framework of the parallel hybrid annotation system......................... 49
Figure 2.13: Concept extraction results (a) Exact matching; (b) Relaxed matching. 52
Figure 2.14: Extraction performance comparison between three different concept types ((a) F-score of exact matching; (b) F-score of relaxed matching) ............... 55
Figure 3.1: Distribution of records of each disease without/with COPD..............64
Figure 3.2: Charlson comorbidity index of comorbidities with/without COPD. ......68
Figure 3.3: Distribution of COPD-relevant articles over COPD-focussed journals. 70
Figure 3.4: Our semi-automatic annotation workflow established in Argo. ..........83
Figure 3.5: GUI of the Manual Annotation Editor component ((a) Segment of one annotated biomedical article; (b) Annotations of peripheral muscle dysfunction (as Problem), muscle dysfunction (as Disease) and muscle (as NamedEntity);(c) Segment of one annotated clinical record; (d) Annotations of ACE inhibitor (as Treatment and Drug)) ................................................................................................ 86
Figure 3.6: F-score measurements of exact and relaxed matching........................ 86
Figure 3.7: Comparative evaluation results of 10-fold cross validation using concept recognisers trained on our corpus.................................................................91
Figure 3.8: Comparative evaluation results of unseen data annotation using concept recognisers trained on our corpus.................................................................95
Figure 4.1: Screenshots of the example of the search results of elevated cholesterol and blood pressure is low through UTS Metathesaurus Browser (a) Hypercholesterolemia is returned for elevated cholesterol; (b) Hypotension is returned for blood pressure is low.................................................................107
Figure 4.2: Screenshots of the example of the search results of high WBC through UTS Metathesaurus Browser (a) High WBC cannot be found in the terminology; (b) High is returned for high; (c) White Blood Cell Count procedure is returned for WBC.
.................................................................................................................................108
Figure 4.3: Example of the formation of Greek phenotypes. ...............................111
Figure 4.4: Example of the formation of English phenotypes............................112
Figure 4.5: PASes performed by Enju (a) C1 (cholesterol elevation); (b) C2 (elevation of cholesterol); (c) C3 (elevated cholesterol); (d) C4 (elevating cholesterol). .............................................................................................................118
Figure 4.6: Nesting PASs of phenotypes (a) Reversible airflow limitation; (b) Management of stable COPD; (c) Elevated arterial carbon dioxide. ..............................119
Figure 4.7: The syntactic transformation of cholesterol elevation..........................120
Figure 4.8: The conversion process between pairs of PASes (a) C1 to C2; (b) C1 to C3; (c) C1 to C4. .....................................................................................................121
Figure 4.9: Screenshots of examples of subcategorisation frames (SubCat Frame) with the probability scores of prepositions in the BioLexicon (a) Results of `elevate` (ARG1#ARG2#PP-in); (b) Results of `decrease` (ARG1#ARG2#PP-in); (c) Results of `treat` (ARG1#ARG2#PP-with); (d) Results of `expose` (ARG1#ARG2#PP-to). 122

Figure 4.10: Normalisation performance using individual methods (a) Clinical domain; (b) Biomedical domain. 130

Figure 4.11: Normalisation performance of hybrid methods in clinical domain (a) Displayed with y axis from 0 to 1; (b) In order to show the differences more clearly, the same diagram is enlarged by using y axis from 0.65 to 0.75. 133

Figure 4.12: Normalisation performance of hybrid methods in biomedical domain (a) Displayed with y axis from 0 to 1; (b) In order to show the differences more clearly, the same diagram is enlarged by using y axis from 0.65 to 1. 134

Figure 4.13: Visualisation of evaluation of all our normalisation methods. 111

Figure 4.14: Examples of the improvement of automatic concept recognition in MeSH (dark grey), the SNOMED-CT (light grey), and both (white). 139
Abstract

LINKING ELECTRONIC HEALTH RECORDS WITH
THE BIOMEDICAL LITERATURE
Xiao Fu
A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2016

Clinical records and biomedical literature, which have grown exponentially in recent years, are important for clinicians to provide personalised treatments and for individual patients to understand their health conditions well. However, essential information is often expressed in natural language. Those expressions in biomedical and clinical domains are distinct, making their processing a daunting task for automated systems. This thesis is the first comprehensive study focussing on concept extraction and multi-level normalisation across biomedical and clinical domains. In this research, we describe our work on 1) developing machine learning-based methods to recognise phenotypic concepts from biomedical and clinical articles; 2) analysing the characteristics of concepts in these two heterogeneous domains and based on this analysis, 3) proposing a normalisation method for linking phenotypic mentions from clinical records and biomedical literature to terminology standards.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
Acknowledgements

I would like to express my sincere gratitude to Prof. Sophia Ananiadou, not only for the guidance and support of my research as my supervisor, but also for her understanding and encouragement like family. Until now, I can still clearly remember the first time we met each other in my hometown in China, and always believed that the coincidence was one of the most important moments in my whole life.

I am extremely thankful to the University of Manchester who provided the financial support for my research and my entire journey in the United Kingdom.

Pursuing a PhD in a foreign country would have been much more difficult for me without the kind help from my colleagues at the National Centre for Text Mining. My immense thanks go to Nhund, Paul, George, and Noha, who have always been there for me. I would like to especially thank my collaborator, Riza, who has generously and constantly given me support for my research and my life like a big sister.

Lastly, words cannot describe how grateful I am to my ever-supportive parents who always respected my decisions and encouraged me to pursue my dreams.
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAD</td>
<td>Acronym and Abbreviation Disambiguation</td>
</tr>
<tr>
<td>AD</td>
<td>Abbreviation Disambiguation</td>
</tr>
<tr>
<td>AHRQ</td>
<td>Agency for Healthcare Research and Quality</td>
</tr>
<tr>
<td>AIDS</td>
<td>Acquired Immune Deficiency Syndrome</td>
</tr>
<tr>
<td>AJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BILOU</td>
<td>Begin/Inside/Last/Outside/Unit-length</td>
</tr>
<tr>
<td>BIO</td>
<td>Begin/Inside/Outside</td>
</tr>
<tr>
<td>CDR</td>
<td>Chemical Disease Relation</td>
</tr>
<tr>
<td>CHF</td>
<td>Chronic Heart Failure</td>
</tr>
<tr>
<td>CLEF</td>
<td>Clinical e-Science Framework</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic Obstructive Pulmonary Disease</td>
</tr>
<tr>
<td>CRFs</td>
<td>Conditional Random Fields</td>
</tr>
<tr>
<td>CUI</td>
<td>Concept Unique Identifier</td>
</tr>
<tr>
<td>cTAKES</td>
<td>Clinical Text Analysis and Knowledge Extraction System</td>
</tr>
<tr>
<td>DA</td>
<td>Domain Adaptation</td>
</tr>
<tr>
<td>DS</td>
<td>Distributional Similarity</td>
</tr>
<tr>
<td>E2G</td>
<td>English to Greek</td>
</tr>
<tr>
<td>Abbr</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>EBM</td>
<td>Evidence-based Medicine</td>
</tr>
<tr>
<td>EMR</td>
<td>Electronic Medical Record</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>G2E</td>
<td>Greek to English</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>HBDA</td>
<td>Hierarchical Bayesian Domain Adaptation</td>
</tr>
<tr>
<td>HCT</td>
<td>Hematocrit</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>HPO</td>
<td>Human Phenotype Ontology</td>
</tr>
<tr>
<td>HPDG</td>
<td>Head-driven Phrase Structure Grammar</td>
</tr>
<tr>
<td>i2b2</td>
<td>Informatics for Integrating Biology and the Bedside</td>
</tr>
<tr>
<td>IAR</td>
<td>Interactive Annotation Refinement</td>
</tr>
<tr>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>International Classification of Diseases-Clinical Modification, Ninth Revision, Clinical Modification</td>
</tr>
<tr>
<td>ICU</td>
<td>Intensive Care Unit</td>
</tr>
<tr>
<td>IE</td>
<td>information Extraction</td>
</tr>
<tr>
<td>LOINt</td>
<td>Logical Observation Identifiers Names and Codes</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
</tr>
<tr>
<td>MK</td>
<td>Meta-Knowledge</td>
</tr>
<tr>
<td>MIMIC II</td>
<td>Multiparameter Intelligent Monitoring in Intensive Care II</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>N</td>
<td>Noun</td>
</tr>
<tr>
<td>NER</td>
<td>Named Entity Recognition</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NICE</td>
<td>National Institute for Health and Care Excellence</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>MMTx</td>
<td>MetaMap Transfer</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>P2S</td>
<td>Plural to Singular</td>
</tr>
<tr>
<td>PASes</td>
<td>Predicate-argument Structures</td>
</tr>
<tr>
<td>PATO</td>
<td>Phenotype and Trait Ontology</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PHI</td>
<td>Protected Health Information</td>
</tr>
<tr>
<td>PMC</td>
<td>PubMed Central</td>
</tr>
<tr>
<td>POS</td>
<td>Part-of-Speech</td>
</tr>
<tr>
<td>PMI</td>
<td>Pointwise Mutual Information</td>
</tr>
<tr>
<td>PPMI</td>
<td>Positive Pointwise Mutual Information</td>
</tr>
<tr>
<td>QE</td>
<td>Query Expansion</td>
</tr>
<tr>
<td>SCL</td>
<td>Structural Correspondence Learning</td>
</tr>
<tr>
<td>SN</td>
<td>Syntactic Normalisation</td>
</tr>
<tr>
<td>SNOMED-CT</td>
<td>Systematised Nomenclature of Medicine-Clinical Terminology</td>
</tr>
<tr>
<td>SS</td>
<td>Synonym Searching</td>
</tr>
<tr>
<td>SubCat Frame</td>
<td>Subcategorisation Frame</td>
</tr>
<tr>
<td>SVMs</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>UBERON</td>
<td>Uber Anatomy Ontology</td>
</tr>
<tr>
<td>UIMA</td>
<td>Unstructured Information Management Architecture</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>UMLS</td>
<td>Unified Medical Language System</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USNLM</td>
<td>United States National Library of Medicine</td>
</tr>
<tr>
<td>UTS</td>
<td>UMLS Terminology Services</td>
</tr>
<tr>
<td>V</td>
<td>Verb</td>
</tr>
<tr>
<td>VVD</td>
<td>Past Tense</td>
</tr>
<tr>
<td>VVG</td>
<td>Gerund/Participle</td>
</tr>
<tr>
<td>WBC</td>
<td>White Blood Cell</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>XMI</td>
<td>XML Metadata Interchange</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This chapter provides an introduction to our research. It begins with the benefits and the challenge of extracting useful information from biomedical and clinical documents, and is followed by an outline of our research hypotheses, objectives and contributions.

1.1 Motivation

Biomedical and clinical knowledge resources in electronic form, including electronic health records and biomedical literature, have grown exponentially in recent years.

The electronic health record is the primary approach to record clinical phenotypes (Roque et al. 2011). The use of an electronic clinical record system, i.e., a computer application that keeps track of the health and medical history of individual patients (Jain et al., 2010; Luo, 2006), has been shown to reduce costs, improve clinical efficiency, as well as improve the quality and standardisation of healthcare provided (Holroyd-Leduc et al. 2011; Yamada 2008). However, in the clinical records such as patient discharge summaries or progression notes,
phenotypic information is often expressed in free-text form which is a convenient means of describing concepts or events but is difficult for manually searching through (Meystre and Haug 2006). The same situation is also found in biomedical literature which conveys detailed scientific findings which can provide evidence to support clinicians in the decision making process (Rosenberg and Donald 1995).

Moreover, the phenotypes in biomedical and clinical documents are quite different as clinical texts consist primarily of brief narrative descriptions (e.g., cholesterol elevation, could not breathe, or blood pressure is high), and acronyms and abbreviations (e.g., high WBC, SOB, or HTN), while biomedical findings are more academic and technical, and mostly expressed in professional medical sublanguage derived from Greek language (e.g., hypercholesterolemia, dyspnea, hypertension, or leukocytosis) (Carroll et al. 2012; Wu and Liu 2011). The performance of linking phenotypic mentions in clinical records to biomedical literature would deteriorate without any normalisation.

Therefore, the increasing availability of biomedical and clinical information has not made it easier for clinicians to filter and receive useful knowledge and incorporate evidence to day-to-day practice, and also impede patients’ attempts to understand their health conditions. In this research, we focus on the development of methods for automatically extracting and normalising phenotypes hidden in biomedical and clinical articles for better health care quality. To the best of our knowledge, ours is the first effort to investigate concept extraction and normalisation in both biomedical and clinical domains.

1.2 Research hypotheses and objectives

To address the problems mentioned above, we raised the following research questions:

- **Q₁** What are the existing approaches for phenotype extraction and annotation?
- **Q₂** What improvements can we introduce to the current methods to achieve better performance for phenotype extraction?
- **Q₃** What are the state-of-the-art approaches for medical corpus annotation?
**Q4** What are the differences between phenotypes in biomedical and clinical domains?

**Q5** Which normalisation approaches can be applied to those two heterogeneous domains?

**Q6** Does the normalisation improve the performance of linking clinical records with biomedical literature?

Before starting our research, the following research hypotheses were proposed:

**H1** Phenotypes in both biomedical and clinical domains can be recognised using existing concept extraction methods and the performance can be improved;

**H2** Even though phenotypes in biomedical and clinical documents are distinct, they can be normalised and identified.

On the basis of these hypotheses, we established the following research objectives:

**O1** To conduct a comprehensive review of existing annotated corpora and annotation approaches for phenotypic named entity recognition (NER);

**O2** To develop phenotypic named entity recognisers to extract phenotypes in biomedical and clinical articles;

**O3** To develop a phenotype annotation workflow with the integration of various text mining techniques;

**O4** To analyse the differences between the phenotypes extracted from biomedical and clinical domains;

**O5** To conduct a comprehensive review of existing techniques for concept normalisation;

**O6** Based on the analysis fulfilled by O4 and O5, to develop implementations of different normalisation methods and observe their impact on linking performance.
How these objectives were achieved is described in the following chapters. Chapter 2 focuses on the fulfilment of objectives $O_1$ and $O_2$ whilst Chapter 3 provides details on how the objective $O_3$ was accomplished. Lastly, we show in Chapter 4 how we reached objectives $O_4$, $O_5$ and $O_6$.

### 1.3 Overview of contributions

Our project, comprising four components/steps, is illustrated in Figure 1.1. Unlabelled biomedical literature and clinical records are firstly collected. With the help of a semi-automatic text mining-assisted annotation tool, all the phenotypes extracted by several embedded machine learning-based named entity recognisers are validated by our domain experts through a graphical user interface (GUI). The resulting annotated corpus of biomedical and clinical articles will then be used to develop the phenotype normalisation methods for improving the linking between biomedical and clinical domain.

The contributions of this research are summarised in the following points:

- **C$_1$** Development of machine learning-based phenotypic named entity recognisers;
- **C$_2$** Development of a text mining-assisted workflow for phenotype annotation and curation;
- **C$_3$** Development of the corpus of biomedical and clinical articles annotated with phenotypic information;
- **C$_4$** Implementations of different normalisation methods based on the analysis of phenotype characteristics;
- **C$_5$** The first comprehensive study on phenotype normalisation in the biomedical and the clinical domains;
- **C$_6$** Improvements of linking phenotypic mentions in clinical and biomedical texts to terminology standards.
Figure 1.1: Framework of linking clinical records with biomedical literature.
Chapter 2

Phenotype extraction

Essential information relevant to medical problems, tests, and treatments is often expressed in biomedical and clinical articles with natural language, making their processing a daunting task for automated systems. One of the steps towards alleviating this problem is concept extraction. In this chapter, we propose a machine learning-based NER system to extract phenotypic concepts without the need for any external knowledge resources to obtain better recognition performances. Three pre- and post-processing methods are investigated, the individual annotation results of which are combined into a final annotation set using two refinement schemes.

2.1 Literature review

Phenotypic concept or phenotype, such as expressions pertaining to drug/therapies (Li et al. 2013; Torii et al. 2011), and procedures/tests (Xu et al. 2010), classically refers to any observable manifestation of an organism, as opposed to genotype, which denotes inheritable information (Blamire 2000). Up
until now, multiple disease characteristics have been termed phenotypes and ideally, individuals sharing a unique phenotype would also ultimately be determined to have a similar underlying biologic or physiologic mechanism(s) to guide the development of therapy where possible (Han et al. 2010).

Phenotype extraction is a critical component of information extraction (IE) (Torii et al. 2011; Uzuner et al. 2010) and has been widely applied to identify obscured and inaccessible information from free-text clinical narratives (Denny et al. 2010; Jonnalagadda et al. 2012) and biomedical documents (Demner-Fushman et al. 2010; Ohno-Machado et al. 2013).

### 2.1.1 Pre-processing

Before extracting phenotypes from natural-language texts, the basic pre-processing procedure consists of the following steps:

- **Sentence segmentation**
  Sentence boundary detector mainly relies on a precompiled lexicon (Patrick and Li 2010), which stores the abbreviations embedded with a period. If the current token is in that lexicon of abbreviations, then it does not end a sentence. But if the next token is capitalized, then the period ends a sentence, that is, the period that ends the abbreviation also ends the sentence. Otherwise, if the token is not an abbreviation and there was a period, exclamation, or question mark at the end of it, and meanwhile the next token was capitalized, this punctuation ended a sentence. In this way, the records will be decomposed into sentences.

- **Tokenisation**
  A white space and punctuation symbol tokeniser (Patrick and Li 2010; Yang 2010) can be applied to separate the sentences into tokens, e.g., words, phrases, symbols, or other meaningful elements.
• **Part-of-speech (POS) tagging**

POS tagging can be performed by the GENIA tagger (Tsuruoka 2005), the Stanford POS tagger (M. Jiang et al. 2011) or the MedPOST tagger (Toutanova et al. 2003) to read text and assign parts of speech to each token, such as noun, verb, adjective, or adverbs, based on its definition and context.

• **Chunking**

A Begin/Inside/Outside (BIO) chunking representation (Yang 2010) is the most common method to annotate and classify individual tokens to make reading and understanding faster and easier, in which each word is assigned into any of the following labels: B (i.e., Beginning of an entity), I (i.e., Inside an entity), and O (i.e., Outside of an entity). Novel chunkers can be made to achieve higher efficiency. For example, the meaningful n-grams for the “Outside” tag as well as three tags (i.e., “Test”, “Treatment”, and “Problem”) were marked in de Bruijn et al. (2011)’s work, in which a better performance was achieved when they replace the tag for “Outside” (e.g., “<Test> <Outside> <Problem>” or “<Treatment> <Outside> <Problem>”) with the lexical item it tags, such as “for” or “of” (e.g., “<Test> <for> <Problem>” or “<Treatment> <of> <Problem>”). In accordance with different objectives, more specific chunkers can also be created, such as for B-test, B-problem, B-treatment, I-test, I-problem, I-treatment, or Outside (Toutanova et al. 2003).

Similar but more detailed than BIO, BILOU (Borthwick 1999; Ratinov and Roth 2009) encode the Beginning (B), the Inside (I) and the Last token (L) of multi-token chunks and differentiate them from Unit-length chunks (U).

• **Feature selection**

Various types of features can then be derived from the training texts and the external resources (de Bruijn et al. 2011; Deléger et al. 2010) to construct a high-dimensional feature space for machine learning-based methods.
Text-oriented features could include token, context, sentence, section and document information (de Bruijn et al. 2011; Toutanova et al. 2003), and external terminology resources of syntactic and semantic tagging, such as the Unified Medical Language System (UMLS) (Bodenreider 2004), the Clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al. 2010) as well as the lists of biomedical abbreviations, have also been demonstrated to have a beneficial impact on extraction performance (Deléger et al. 2010). Medical Language Extraction and Encoding System (MedLEE) (Friedman 1997; Friedman et al. 2001), KnowledgeMap (Denny et al. 2005), MGREP concept mapping engine (Bhatia et al. 2009), MetaMap (Aronson et al. 2000) and ConText (de Bruijn et al. 2011) can be utilised to map terms or concepts from those external natural language processing systems.

2.1.2 Post-processing

Some post-processing methods have also been applied to improve the extraction performance.

- **Lexicon filtering**
  A context-based filter-out step can be employed to exclude the mentions that belong to a general English word list (Deléger et al. 2010) or some specific situations (Yang 2010).

- **Identify unseen words**
  Previously unknown words in the test set could be accurately labelled by techniques like the Brown clustering algorithm (Brown et al. 1992), which groups semantic concepts and POS tags into clusters based on the contexts in which they occur (de Bruijn et al. 2011).
- **Identify typing errors**

  The self-developed lists of misspellings or the spell check functions (Patrick and Li 2010) can also be used to avoid the errors caused by mistyping.

### 2.1.3 Phenotype extraction in the clinical domain

In recent years, research has been conducted to apply rule- and/or machine learning-based approaches to clinical records to assist scientists in the extraction of valuable information. A discriminative semi-Markov hidden Markov model (HMM) was proposed by de Bruijn et al. (2011) to recognise clinical entities (i.e., Problem, Treatment and Test) within discharge summaries and progress reports. A wide range of textual features generated from both training texts and external medical knowledge bases (i.e., UMLS, cTAKES, and MEDLINE (MEDLINE 2006)) for semantic and syntactic tagging were employed and demonstrated to have a beneficial influence on the model performance.

Jiang et al. (2011) developed a hybrid clinical concept extraction system. In combination with rule-based post-processing methods, two machine learning algorithms, i.e., conditional random fields (CRFs) and support vector machines (SVMs) were separately applied to construct the named entity recogniser, and the performances of these two methods were evaluated and compared. CRFs were observed to perform better than SVMs, and the performance of clinical named entity recognisers was significantly enhanced by the addition of semantic features derived from existing external knowledge sources.

Similarly, a hybrid model based on a cascaded approach for medication extraction was built by Patrick and Li (Patrick and Li 2010). CRF-based and SVM-based classifiers and pattern matching rules were incorporated. Since only 17 annotated clinical records were initially provided which are too little to train the models, another 145 records were manually annotated by themselves to augment the training set. The performance of their model indicates that as more gold standard annotations become available, the better performance of automatic concept recognisers would be achieved.
Pedersen (2006) reported some success in using supervised (SenseTools) and unsupervised methods (SenseClusters) with the lexical features, i.e., unigrams, bigrams, and trigrams, on medical records to identify the smoker-status of the patients. Their results also indicated that the supervised methods were much more effective than unsupervised ones if enough manually annotated data was available. Jain et al. (2010) proposed that semantic query expansion (QE) techniques, which involve evaluating a user's input and then suggesting related queries to match additional documents, can be utilised to improve retrieval performance, evaluating on nursing notes drawn from the electronic medical record (EMR) system.

Byrd et al. (2013) introduced a hybrid natural language processing (NLP) tool for improving early recognition and diagnosis of chronic heart failure (CHF). Interactive annotation refinement (IAR) was used to create rule-based entity extractors, and CHF signs and symptoms were extracted from the clinical records of primary care patients according to the Framingham CHF criteria (McKee et al. 1971). Kang et al. (2010) integrated six named entity recognisers and chunkers, namely, ABNER (Settles 2015), Lingpipe (Alias-i 2008), OpenNLP Chunker and OpenNLPNer (OpenNLP 2008), Peregrine (Schuemie et al. 2007), and StanfordNer (Finkel et al. 2005), to annotate clinical records. In order to generate composite annotations, majority voting (Ruta and Gabrys 2005) which considers an annotation as valid if results from any three of the six systems are identical, was then applied. A combined annotation system for clinical records was proven by their results to substantially outperform any of the individual systems. It is not possible to compare precisely the performance of especially machine-learning based systems where different training data has been used.

### 2.1.4 Phenotype extraction in the biomedical domain

Biomedical literature has also grown exponentially in recent years, and NLP researchers often face a concomitant problem to make huge amounts of information in publications more accessible and useful for scientists. In order to alleviate and solve this problem, much work has been done to apply text mining techniques to
biomedical texts, such as the CRAB text mining tool (Korhonen et al. 2012) used to support cancer risk assessment, the EventMine-MK (Meta-Knowledge) system (Miwa et al. 2012) to extract semantically enriched bio-events, and the search application ASCOT that aims to address information overload problem encountered in clinical trials as well as to assist the creation of new protocols (Korkontzelos et al. 2012).

### 2.1.5 Evidence-based medicine

Evidence-based medicine (EBM) is defined as “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients”, and the practice of EBM means linking personalised clinical information in the EMR with the best available external scientific findings from systematic research (Sackett et al. 1996; Sim et al. 2001). The current studies about EBM are mainly focusing on mapping clinical reports and biomedical texts to structured knowledge resources (Aronson and Lang 2010b), also called standardised terminological systems (Lieberman et al. 2005; Saitwal et al. 2012), such as UMLS Metathesaurus and Systematised Nomenclature of Medicine-Clinical Terminology (SNOMED-CT) (SNOMED-CT 2014), to find the relationship between them.

UMLS Metathesaurus developed by the United States National Library of Medicine (USNLM) is a set of codes, thesauri, classifications and controlled terms in biomedical domain (Aronson 2001; USNLM 2013), and this knowledge source has proved useful for providing machine-readable information that can be used to underpin the semantic interoperability of different terminological systems by linking terms and concepts with the same meaning (Saitwal et al. 2012). This metathesaurus contains the codes from Medical Subject Headings (MeSH) (Coletti and Bleich 2001), the International Classification of Diseases (ICD) (WHO 2014a), RxNorm (Liu et al. 2005), Logical Observation Identifiers Names and Codes (LOINC) (Forrey et al. 1996), and SNOMED-CT.

SNOMED-CT is a comprehensive, multilingual clinical healthcare terminology owned, maintained and distributed by the International Health Terminology
Standards Development Organisation (IHTSDO) (SNOMED-CT 2014). It makes a significant contribution towards improving the quality and safety of healthcare by enabling the exchange of electronic health information (IHTSDO 2014). Nevertheless, a number of classification errors in SNOMED-CT has been recently noticed by researchers (Nadkarni 2010). There has been a wide range of applications of those knowledge sources. Brennan and Aronson (2003) introduced a method to link emails of patients to nurses with complex health knowledge resources to facilitate patient access to health information.

There are also some other medication terminological and coding systems, like MedEX (Xu et al. 2010) which focuses on identifying medication information, including drug names, signature information (frequency, dosage, strength, etc.) and contextual level information (status, temporal information, etc.) from clinical notes, and MedLEE, whose goal is to extract, structure, and encode clinical information in textual patient reports. McCormick et al. (2008) observed that the performance of a classifier for patient's smoking status stratification based on semantic features generated by the MedLEE parser was better than that using purely lexical features. In addition, Saitwal et al. (2012) proposed a cross-terminology mapping method to alleviate problems caused by the heterogeneity (Torii et al. 2011) of clinical data. Automatic, semi-automatic and manual mapping methods were employed to map drug codes from electronic health records to SNOMED CT and the UMLS Metathesaurus.

The MetaMap Indexing method which is a dictionary-based system offering a connection between the biomedical texts and the UMLS Metathesaurus concepts, was used to identify and rank UMLS concepts in e-mail; while the problem was the free-text messages of patients who are the lay people did not contain many UMLS concepts as they prefer to use idioms and colloquial words. Mendonça et al. (2002) described the use of the EMR to determine the relevance of research evidence and thus the literature matching the individual’s medical records would be presented first instead of the usual presentation in reverse chronological order. A medical entities dictionary was applied to map terms in the EMR to UMLS concepts.
However, most of their results did not agree with the selection of clinicians. Based on the research conducted by Kang et al. (2012), in comparison with five statistical-based systems, one dictionary-based system, and the ensemble systems, MetaMap generated the worst results when being applied to extract medical problems, tests, and treatments from clinical records. Therefore, simply mapping to external knowledge sources cannot successfully link biomedical literature with clinical records.

2.2 Dataset

This study is performed on patient discharge summaries and progress notes in the 2010 Informatics for Integrating Biology and the Bedside (i2b2)/VA challenge data set (Uzuner et al. 2011). Seventy-three (73) human annotated records were used for system training, while the test data set was comprised of 256 annotated records. This split of the training and the test sets was decided by the i2b2 organisers.

For each record, there are two types of files. One is the report file, which has already been split into sentences, and the other is the annotation file, in which annotations are specified by means of line and word numbers that indicate text spans corresponding to concepts. Table 2.1 shows an example of a sentence “Trauma series demonstrated no evidence of a pelvic fracture.” in report files and its corresponding annotations in annotation files, where t and c denote type and concept, respectively. The annotations mean that the concept trauma series belongs to the category test. 44 is the sentence number, whilst 0 and 1 are the word numbers of trauma and series in the sentence. In the training set, there are 3,711 entities which were assigned the problem label, 2,620 assigned the test label and 2,867 annotated as treatment, while in the test set, 13,068, 9,641 and 9,550 entities were annotated as problem, test and treatment, respectively.
Table 2.1: Examples from a pair of report and annotation files.

<table>
<thead>
<tr>
<th>Sentences (Report file)</th>
<th>Concept annotations (Annotation file)</th>
</tr>
</thead>
</table>
| *Trauma series demonstrated no evidence of a pelvic fracture.* | c='trauma series’ 44:0 44:1||t='test’  
c='a pelvic fracture’ 44:6 44:8||t='problem’ |

2.3 Methodology

We proposed a machine learning-based named entity recognition system to extract clinical concepts without using any external knowledge resources. Three pre- and post-processing methods were investigated, the individual annotation results of which were then combined into a final annotation set using two refinement schemes.

2.3.1 Baseline

We selected NERsuite\(^1\), which is a freely available NER tagger based on the CRFsuite which is an implementation of conditional random fields (CRFs) (Okazaki et al. 2010), to generate the concept annotations.

- **CRFs**

  CRFs are a type of discriminative undirected probabilistic graphical model, which is used to encode known relationships between observations and construct consistent interpretations. It is defined as follows (Lafferty et al. 2001):

  Let \( G = (V, E) \) be a graph such that \( Y = (Y_v)_{v \in V} \), so that \( Y \) is indexed by the vertices of \( G \). Then \((X, Y)\) is a conditional random field in case, when

\(^1\) http://nersuite.nlplab.org/
conditioned on \( X \), the random variables \( Y_v \) obey the Markov property with respect to the graph: 
\[
p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v),
\]
where \( w \sim v \) means that \( w \) and \( v \) are neighbours in \( G \).

Hence, a CRF is an undirected graphical model based on two jointly distributed random variables, namely, observations \( X \) and random variables \( Y \) to model the conditional distribution \( p(Y | X) \).

NERsuite consists of three modules, a tokeniser, a modified version of GENIA tagger, and a NER. The procedure is as shown in Figure 2.1: First, the sentence-split report files were processed by the tokeniser which was used to segment each sentence into tokens, and compute the position of each token in the sentence. The modified GENIA tagger was then applied to produce three token features, i.e., the POS tags, lemmas, and chunk tags. A NER model was then trained to assign the label to each token in the test corpus.

![Figure 2.1: Framework of baseline.](image-url)
The required features for training the NER model are:

1) correct named entity label of the token
2) the byte position of the first letter of the token
3) the byte position one past the last letter of the token
4) raw token (word)
5) lemma
6) POS tag
7) chunk tag
8) any other attributes, e.g. dictionary features

However, only the concept types instead of named entity labels of each token are provided in the annotated files (as shown in Figure 2.2 (a)). Thus, in order to train the NER model, the annotated records need to be converted into a BIO format to obtain the correct named entity label for individual tokens. Consequently, we used a total of seven possible named entity labels, namely, “B-problem, I-problem, B-test, I-test, B-treatment, I-treatment, and O”. The transformed annotations are shown in Figure 2.2 (b) in the first column.

For example, known allergies (Problem) is transformed to known (B-problem), and allergies (I-problem), while the named entity label of drugs, i.e., Treatment, is transformed to B-treatment. The sentence number and the token position are converted to byte position as well. Moreover, the example sentence in Table 2.1 will also be transformed to the token-level annotation shown in Table 2.2.
(a)

c="drugs" 12:8 12:8||t="treatment"
c="known allergies" 12:5 12:6||t="problem"

(b)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>0</td>
<td>7</td>
<td>Patient</td>
<td>Patient</td>
<td>NN</td>
</tr>
<tr>
<td>O</td>
<td>8</td>
<td>15</td>
<td>recorded</td>
<td>record</td>
<td>VBD</td>
</tr>
<tr>
<td>O</td>
<td>17</td>
<td>19</td>
<td>as</td>
<td>as</td>
<td>IN</td>
</tr>
<tr>
<td>O</td>
<td>20</td>
<td>26</td>
<td>having</td>
<td>have</td>
<td>VBG</td>
</tr>
<tr>
<td>B-problem</td>
<td>27</td>
<td>32</td>
<td>Known</td>
<td>Known</td>
<td>NNP</td>
</tr>
<tr>
<td>I-problem</td>
<td>33</td>
<td>42</td>
<td>Allergies</td>
<td>Allergy</td>
<td>NNPS</td>
</tr>
<tr>
<td>O</td>
<td>43</td>
<td>45</td>
<td>to</td>
<td>to</td>
<td>TO</td>
</tr>
<tr>
<td>B-treatment</td>
<td>46</td>
<td>51</td>
<td>Drugs</td>
<td>Drug</td>
<td>NNS</td>
</tr>
</tbody>
</table>

Figure 2.2: Example snippets of

(a) Annotation files from the i2b2/VA challenge;

(b) Converted named entity label (the first column) and byte positions (the second and the third columns) for training the NERsuite model.
Table 2.2: Transformed annotations of “Trauma series demonstrated no evidence of a pelvic fracture.”.

<table>
<thead>
<tr>
<th>Named entity label</th>
<th>Beginning position</th>
<th>Past-the-end position</th>
<th>Token</th>
<th>Lemma</th>
<th>POS</th>
<th>Chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-test</td>
<td>0</td>
<td>6</td>
<td>Trauma</td>
<td>Trauma</td>
<td>NN</td>
<td>B-NP</td>
</tr>
<tr>
<td>I-test</td>
<td>7</td>
<td>13</td>
<td>series</td>
<td>series</td>
<td>NN</td>
<td>I-NP</td>
</tr>
<tr>
<td>O</td>
<td>14</td>
<td>26</td>
<td>demonstrated</td>
<td>demonstrate</td>
<td>VBD</td>
<td>B-VP</td>
</tr>
<tr>
<td>O</td>
<td>27</td>
<td>29</td>
<td>no</td>
<td>no</td>
<td>DT</td>
<td>B-NP</td>
</tr>
<tr>
<td>O</td>
<td>30</td>
<td>38</td>
<td>evidence</td>
<td>evidence</td>
<td>NN</td>
<td>I-NP</td>
</tr>
<tr>
<td>O</td>
<td>39</td>
<td>41</td>
<td>of</td>
<td>of</td>
<td>IN</td>
<td>B-PP</td>
</tr>
<tr>
<td>B-problem</td>
<td>42</td>
<td>43</td>
<td>a</td>
<td>a</td>
<td>DT</td>
<td>B-NP</td>
</tr>
<tr>
<td>I-problem</td>
<td>44</td>
<td>50</td>
<td>pelvic</td>
<td>pelvic</td>
<td>JJ</td>
<td>I-NP</td>
</tr>
<tr>
<td>I-problem</td>
<td>51</td>
<td>59</td>
<td>fracture</td>
<td>fracture</td>
<td>NN</td>
<td>I-NP</td>
</tr>
<tr>
<td>O</td>
<td>60</td>
<td>61</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>O</td>
</tr>
</tbody>
</table>
The resulting files with the named entity labels are in the format shown in Figure 2.3, in which the predicted labels in the BIO format are displayed in the last (i.e., the seventh) column.

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>67</td>
<td>There</td>
<td>There</td>
<td>EX</td>
<td>B-NP</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>68</td>
<td>70</td>
<td>is</td>
<td>be</td>
<td>VBZ</td>
<td>B-VP</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>71</td>
<td>75</td>
<td>mild</td>
<td>mild</td>
<td>JJ</td>
<td>B-NP</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>76</td>
<td>82</td>
<td>mitral</td>
<td>mitral</td>
<td>JJ</td>
<td>I-NP</td>
<td>B-problem</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>90</td>
<td>annular</td>
<td>annular</td>
<td>JJ</td>
<td>I-NP</td>
<td>I-problem</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>104</td>
<td>calcification</td>
<td>calcification</td>
<td>NN</td>
<td>I-NP</td>
<td>I-problem</td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>105</td>
<td>.</td>
<td>.</td>
<td></td>
<td>O</td>
<td>O</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.3: Example snippet of test files annotated by NERsuite

### 2.3.2 Truecasing

While capitalisation is typically used only to begin sentences, we have observed that in these clinical documents several full words are also capitalised for the purpose of emphasis. Given these considerations, we considered the application of the truecasing method, generally used to restore the correct case of tokens in raw texts to consistently transform token expressions to their canonical forms (Lita et al. 2003; Pyysalo and Ananiadou 2014).

The true case of a token is predicted by the most likely case of that token. It will transform 1) the sentence-initial words (e.g., *No driving for one month*), 2) words in title case (e.g., *Patient May Shower, No Baths*, or the titles of sections, tables, etc.), and 3) words in all-upper case contexts (e.g., *ASPIRIN 81 MG DAILY*, or some publication titles.) to lowercase letters, i.e., *no driving for one month*, *patient may shower, no baths*, and *aspirin 81 mg daily*. A model derived from a random sample of the 2012 PubMed baseline distribution was implemented here, and local context was captured through a trigram language model, but the case label will be decided on the most likely meaning at a sentence level. The procedure of truecasing is depicted in Figure 2.4.
Figure 2.4: Framework of truecasing using modules in Argo.

This model has been included in Argo\(^1\) (Rak et al. 2012, 2013) as the Truecase Asciiifier module. The architecture of Argo is shown in Figure 2.5.

---

\(^1\) [http://argo.nactem.ac.uk/](http://argo.nactem.ac.uk/)
Argo complies with the Unstructured Information Management Architecture (UIMA) standard. As a text mining platform, its overarching goal is to support the extraction of structured information from unstructured data, or text. It comes with a library of UIMA-compliant text mining components ranging from data readers and writers to parsers, taggers, and named entity recognisers, which is continuously growing in size thanks to developers who contribute their own tools. Text mining solutions are realised as pipelines or what we call workflows in Argo, and to support solution or workflow developers, Argo has a workflow diagramming interface. It also has an interface for manual annotations, as previously mentioned, that data curators or domain experts can use to add, remove or correct automatically generated annotations. In Figure 2.6, which is a screenshot of the workflow diagramming interface of the GENIA Tagger component, on the left side you will find the library of elementary components. The centre panel acts as the canvas where the elementary components, such as GENIA Tagger, can be arranged and connected to each other to form a meaningful workflow. The right panel that displays details of a selected component can be used by a user to check the compatibility of input and output types.

Figure 2.6: A screenshot of the workflow diagramming interface of Argo.
(Example of the GENIA Tagger component)
The Truecase Asciifier module is shown in Figure 2.7, and it was employed in our work to generate tokens in their normalised case form, and the examples of truecasing features for *Present* and *Illness* are displayed below. Through the Truecase Asciifier module, *Present* and *Illness* in the original texts are converted to their true cases, i.e., *present* and *illness*, respectively.

Figure 2.7: Truecase Asciifier module in Argo.

- **Examples of truecasing features**

  ```xml
  <eupmc:AlternateToken xmi:id="30012" sofa="1" begin="509" end="516" ascii="Present" ascii_truecase="present" token="1605" asciiBase="Present" asciiBase_truecase="present"/>
  <eupmc:AlternateToken xmi:id="30021" sofa="1" begin="517" end="524" ascii="Illness" ascii_truecase="illness" token="1614" asciiBase="Illness" asciiBase_truecase="illness"/>
  ``

Truecasing will be performed before tokenisation (i.e., on input sentences) as it takes into consideration the context surrounding any given token.
2.3.3 Abbreviation disambiguation

Acronyms and abbreviations widely appear in clinical records as well as biomedical literature, some of which follow certain conventions whereas others are ambiguous (Carroll et al. 2012). For instance, *q.d.* is the standard acronym for 'quaie die', which means once a day. *MI*, having more than 80 possible full forms, can stand for *myocardial infarction, mitotic index*, and *myo-inositol*, while the expanded form of *CAD* could be any of *coronary artery disease, computer-aided diagnosis*, and *caldesmon*, etc., depending on the context. Therefore, a disambiguation process is crucial to ensure the correct interpretation of the records (Uzuner et al. 2011). In our study, we examined the impact of abbreviation disambiguation (AD) on NER by transforming the acronyms and abbreviations or short forms in the report files to their suitable expanded full forms estimated by AcroMine Disambiguator, a word sense disambiguation classifier trained on MEDLINE abstracts (Okazaki et al. 2010).

To choose correct expanded forms for each abbreviation *a*, the likelihood of long-form candidates for a candidate *c* is calculated through the Equation 2.1, which is a modified version of the original C-value method (Frantzi et al. 2000).

\[
LH_a(c) = \text{freq}(a,c) - \sum_{t \in c} \frac{\text{freq}(a,t)}{\sum_{w \in c} \text{freq}(a,w)}
\]  

(2.1)

In Equation 2.1, *a* is an abbreviation (i.e., short form) and *c* is a candidate of long form for the abbreviation *a*. The first term of the formula (i.e., \(\text{freq}(a,c)\)) denotes the frequency of occurrence of the candidate *c* with the original abbreviation *a* in the contexts. For the second term, *T_c* is a set of nested candidates of expanded form, each of which consists of a preceding word followed by the candidate *c*. Given a long-form candidate \(t \in T_c\), \(\frac{\text{freq}(a,t)}{\sum_{w \in T_c} \text{freq}(a,w)}\) presents the occurrence probability of candidate *t* in the nested candidate set *T_c*. Therefore, the second term of the formula calculates the weighted average of the frequency of occurrence of nested candidates based on the frequency of candidate *c* (Okazaki and Ananiadou 2006).

---

2 http://www.nactem.ac.uk/software/acromine_disambiguation/
3 http://www.nactem.ac.uk/software/acromine/
The framework of disambiguation of acronyms and abbreviations in our work is displayed in Figure 2.8, and the examples of AcroMine dictionary consisting of the most suitable long-form (or expanded-form) candidates (e.g., *heart rate*, *atrial fibrillation*, and *acetylsalicylic acid*) for short-form terms (*HR*, *AFIB*, and *ASA*) are shown below.

![Diagram of Acromine disambiguation](image)

**Figure 2.8: Framework of abbreviation disambiguation.**

- **Examples of AcroMine dictionary**

  ```xml
  <Acronym type="global" pmid="24314785" location="AbstractText">  
  <ExpandedForm>heart rate</ExpandedForm>  
  <ShortForm begin="3555" end="3557">HR</ShortForm>  
  </Acronym>  
  
  <Acronym type="global" pmid="24314785" location="AbstractText">  
  <ExpandedForm>atrial fibrillation</ExpandedForm>  
  <ShortForm begin="3796" end="3800">AFIB</ShortForm>  
  </Acronym>  
  
  <Acronym type="global" pmid="24314785" location="AbstractText">  
  <ExpandedForm>acetylsalicylic acid</ExpandedForm>  
  <ShortForm begin="662" end="665">ASA</ShortForm>  
  </Acronym>
  ```
2.3.4 Distributional similarity

After we automatically annotated the test corpus using the newly trained NERsuite model, in order to measure distributional similarity (DS) between two words, the words are represented by context vectors and then the distance between the vectors can be calculated. A distributional thesaurus, in which each word is associated with a list of other words according to their DS scores (Carroll et al. 2012), was constructed using the training documents. The thesaurus was then applied to reassign concept types to tokens in the initial results for improving recall, based on the intuition that similar words tend to have the same concept type.

Pair-wise DS scores between words in the documents were computed using Lin’s measurement (Lin 1998), as shown in Equation 2.2.

\[
sim(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}
\]  

(2.2)

In Equation 2.2, \( w \) and \( r \) represent words and the relationship between two words, respectively. \( I(w, r, w') \) is equivalent to the pointwise mutual information (PMI) between \( w \) and \( w' \) (see Equation 2.3).

\[
I(w, r, w') = \log \frac{\|w, r, w'\| \times \|*, r, *\|}{\|w, r, *\| \times \|*, r, w'\|}
\]  

(2.3)

In Equation 2.3, \( \|w, r, w'\| \) denotes the frequency of the triple \((w, r, w')\) in the corpus, and * means there are no specific words. \( T(w) \) means the set of pairs \((r, w')\) such that \( I(w, r, w') \) is positive. Here we used the 4 types of proximity relationships as in the work of Carroll et al. (2012), shown in Table 2.3.

\( I(w, r, w') \) can also be presented by the PMI function (see Equation 2.4.)

\[
I(w, r, w') = \frac{f(w, r, w')}{f(w) \cdot f(w')}
\]  

(2.4)
In Equation 2.4, \( f(w,r,w') = \| w,r,w' \|/\| *,*,* \| \) and \( f(w) = f(w,r,*) \). In addition, since only positive values are considered in \( T(w) \), what is computed in fact amounts to the positive pointwise mutual information (PPMI).

Table 2.3: Proximity relationships used to calculate DS.

<table>
<thead>
<tr>
<th>Relation name</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>prev</td>
<td>Previous word</td>
</tr>
<tr>
<td>prev_window</td>
<td>Word within a distance of 2-5 words to the left</td>
</tr>
<tr>
<td>next</td>
<td>Next word</td>
</tr>
<tr>
<td>next_window</td>
<td>Word within a distance of 2-5 words to the right</td>
</tr>
</tbody>
</table>

The framework is shown in Figure 2.9. Figure 2.10 displays a simple example of how the distributional thesaurus works. Initially, the phenotype nausea was wrongly classified as Test. However, according to the list of words that are most similar to nausea in the distributional thesaurus, in which the numbers in the second column are DS scores and concept types are in the third column, the label of nausea is then correctly reassigned as Problem as most of its similar words are problems.

Figure 2.9: Framework of using distributional similarity.
CHAPTER 2. PHENOTYPE EXTRACTION

She awoke with nausea on the day of admission and pleuritic chest pain.

Distributional similarity

Figure 2.10: Example of how the distributional thesaurus works.

2.3.5 Hybrid methods

Since a hybrid annotation system has been demonstrated to have a better performance than any of the individual systems (Kang et al. 2012; Uzuner et al. 2011), we combined the three aforementioned techniques (i.e., truecasing, AD, and DS) into a final annotation system by two schemes. Each of the schemes is illustrated in Figures 2.11 and 2.12, respectively.

The first is a sequential scheme in which the training and test documents were processed by the truecasing and the AD modules consecutively. In this way, we prevent the AD module from treating words in all uppercase letters as short forms and incorrectly expanding them. For example, AD will typically expand ‘END’ to ‘endurance’, but with the application of the truecasing module, the former will be first transformed to ‘end’, hence avoiding its unnecessary expansion. Next, the texts with the correct case information and expanded forms were used to train and test the NERsuite model to obtain the concept extraction results. A distributional thesaurus built on the new training documents was then employed to assign new annotations to the test documents.

As Kang et al. (2010) indicated that the majority voting method can be used to generate composite annotations when more than one annotation systems are used, in the second scheme, a parallel one, we compute the union of the annotation results of these three systems. If the concept annotations provided by any two of the three systems are identical, they are generated as final, combined annotations. Otherwise,
the result of truecasing is considered as the final annotation, since this technique performed best on the training texts.

### 2.4 Results and discussion

The annotation systems were evaluated using precision, recall and $F_1$-measure (or balanced F-measure), which are calculated as Equation 2.5, 2.6, and 2.7, respectively.

\[
\text{Precision} (P) = \frac{TP}{(TP + FP)} \tag{2.5}
\]

\[
\text{Recall} (R) = \frac{TP}{(TP + FN)} \tag{2.6}
\]

\[
F_1 = \frac{2 \times P \times R}{(P + R)} \tag{2.7}
\]

Where TP, FP, and FN are the numbers of true positives, false positives, and false negatives, respectively. TP, FP and FN are defined in Table 2.4. TP refers to the amount of clinical concepts correctly identified by the annotation system, while FP denotes the number of non-clinical concepts wrongly annotated as clinical concept, and FN represents how many clinical concepts were not recognised by the annotation system. Precision measures the extent to which the mentions recognised by the system were actually correct; recall measures the extent to which the system recognised all of the entity mentions that it was supposed to recognise. $F_1$-measure is the harmonic mean of precision and recall, which provides an overall performance measure for the system.

Table 2.4: Definition of TP, FP, and FN.

<table>
<thead>
<tr>
<th>Condition (Gold standard)</th>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test outcome</strong> (Annotation results)</td>
<td><strong>Positive</strong></td>
<td><strong>TP</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Negative</strong></td>
<td><strong>FN</strong></td>
</tr>
</tbody>
</table>
Figure 2.11: Framework of the sequential hybrid annotation system.
Figure 2.5: Framework of the parallel hybrid annotation system.
Two different modes of matching evaluation were applied:

- **Exact matching**
  This evaluation method considers a system annotation as correct only if it has the same concept type label and exactly the same boundaries as a gold standard annotation.

- **Relaxed matching**
  Relaxed string matching counts even a partially overlapping system annotation as correct as long the non-overlapping tokens consist of only articles and modifiers (i.e., they have only “DT (Determiners)”, “JJ (Adjectives)” or “RB (Adverbs)” as POS tags).

### 2.4.1 Results

We explored the performance of truecasing, AD and DS alone, and various combinations of these three methods for clinical concept annotation. Table 2.5 summarises the performance of our systems on the 256 test records based on exact and relaxed matching. The former requires that automatically generated annotations are exactly the same as those in the gold standard, while the latter additionally takes into account annotations with relaxed boundaries as long as they have overlapping spans (Kang et al. 2012; Uzuner et al. 2011). The same results are visualised in Figure 2.13.

The truecasing method obtained an F-score of 0.7586 according to exact matching, making it our best performing concept annotation system. However, the AD method and the DS method did not show any improvements. The F-score of the sequential combination of the truecasing and AD or DS methods are 0.7556 and 0.6745, respectively, which are worse than that of the purely truecasing-based system and even that of the baseline. The sequential combination of these three methods further reduced the F-score to 0.6877. Nevertheless, integrating those three methods using the parallel scheme achieved better performance than the baseline and the parallel combination of truecasing and AD or DS.

The best F-score measurement using relaxed matching is 0.8444, achieved by the parallel combination of truecasing and AD. The parallel combination of
truecasing, AD, and DS, truecasing on its own, and the sequential combination of truecasing and AD models outperformed the baseline as well.

Table 2.5: Exact and relaxed matching results for concept extraction.
(T, truecasing; S, the sequential scheme; P, the parallel scheme.)

<table>
<thead>
<tr>
<th>Systems</th>
<th>Exact Matching</th>
<th>Relaxed Matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7132</td>
<td>0.8054</td>
</tr>
<tr>
<td>T</td>
<td>0.7147</td>
<td>0.8083</td>
</tr>
<tr>
<td>AD</td>
<td>0.7099</td>
<td>0.8014</td>
</tr>
<tr>
<td>DS</td>
<td>0.6791</td>
<td>0.6554</td>
</tr>
<tr>
<td>T + AD (S)</td>
<td>0.7115</td>
<td>0.8055</td>
</tr>
<tr>
<td>T + DS (S)</td>
<td>0.6865</td>
<td>0.6630</td>
</tr>
<tr>
<td>T + AD + DS (S)</td>
<td>0.7210</td>
<td>0.6573</td>
</tr>
<tr>
<td>T + AD (P)</td>
<td>0.7283</td>
<td>0.7849</td>
</tr>
<tr>
<td>T + DS (P)</td>
<td>0.7002</td>
<td>0.6772</td>
</tr>
<tr>
<td>T + AD + DS (P)</td>
<td>0.8047</td>
<td>0.7140</td>
</tr>
</tbody>
</table>
Figure 2.6: Concept extraction results (a) Exact matching; (b) Relaxed matching.
We also assessed the extraction performance for each entity type (i.e., problem, treatment, and test), as shown in Table 2.6. As shown in Figure 2.14, the comparison of F-scores on different concept types is visualised. Results for the test type indicate the highest F-scores, with around 0.77 using exact matching for all the systems except DS. Based on relaxed matching, the systems excluding the sequential combination of truecasing, AD and DS achieved the best results in problem extraction with F-scores over 0.82.

In our system-level evaluation, truecasing obtained the best performance in recognising problems and treatments on the test records using exact matching, while none of the methods improved extraction performance for tests. The highest F-score using relaxed matching was also achieved by truecasing; adding AD to it employing the parallel combination scheme resulted in the highest improvements in both treatment and test extraction.

Table 2.6: Exact and relaxed matching results for each concept type
(a) Problem; (b) Treatment; (c) Test.

<table>
<thead>
<tr>
<th>Concept type:</th>
<th>Exact matching</th>
<th>Relaxed matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td><strong>Problem</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7094</td>
<td>0.7757</td>
</tr>
<tr>
<td>T</td>
<td>0.7138</td>
<td>0.7803</td>
</tr>
<tr>
<td>AD</td>
<td>0.7071</td>
<td>0.7742</td>
</tr>
<tr>
<td>DS</td>
<td>0.6562</td>
<td>0.7175</td>
</tr>
<tr>
<td>T + AD (S)</td>
<td>0.7102</td>
<td>0.7793</td>
</tr>
<tr>
<td>T + AD + DS (S)</td>
<td>0.7181</td>
<td>0.5596</td>
</tr>
<tr>
<td>T + AD (P)</td>
<td>0.7245</td>
<td>0.7559</td>
</tr>
<tr>
<td>T + AD + DS (P)</td>
<td>0.7099</td>
<td>0.7757</td>
</tr>
</tbody>
</table>

(a)
## Concept type: Treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Exact matching</th>
<th>Relaxed matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.6846</td>
<td>0.8237</td>
</tr>
<tr>
<td>T</td>
<td>0.6890</td>
<td>0.8271</td>
</tr>
<tr>
<td>AD</td>
<td>0.6854</td>
<td>0.8196</td>
</tr>
<tr>
<td>DS</td>
<td>0.6304</td>
<td>0.7527</td>
</tr>
<tr>
<td>T + AD (S)</td>
<td>0.6881</td>
<td>0.8232</td>
</tr>
<tr>
<td>T + AD + DS (S)</td>
<td>0.6974</td>
<td>0.7827</td>
</tr>
<tr>
<td>T + AD (P)</td>
<td>0.7032</td>
<td>0.8063</td>
</tr>
<tr>
<td>T + AD + DS (P)</td>
<td>0.6849</td>
<td>0.8251</td>
</tr>
</tbody>
</table>

## Concept type: Test

<table>
<thead>
<tr>
<th>Test</th>
<th>Exact matching</th>
<th>Relaxed matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall</td>
<td>Precision</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7473</td>
<td>0.8292</td>
</tr>
<tr>
<td>T</td>
<td>0.7493</td>
<td>0.8262</td>
</tr>
<tr>
<td>AD</td>
<td>0.7385</td>
<td>0.8213</td>
</tr>
<tr>
<td>DS</td>
<td>0.5110</td>
<td>0.5685</td>
</tr>
<tr>
<td>T + AD (S)</td>
<td>0.7368</td>
<td>0.8252</td>
</tr>
<tr>
<td>T + AD + DS (S)</td>
<td>0.7487</td>
<td>0.7124</td>
</tr>
<tr>
<td>T + AD (P)</td>
<td>0.7589</td>
<td>0.8051</td>
</tr>
<tr>
<td>T + AD + DS (P)</td>
<td>0.7419</td>
<td>0.8298</td>
</tr>
</tbody>
</table>
Figure 2.7: Extraction performance comparison between three different concept types

(a) F-score of exact matching; (b) F-score of relaxed matching.
2.4.2 Discussion

In our study, several NLP methods were investigated to construct NER systems for clinical concept extraction. Instead of using large-dimensional bags of complex features and rules derived from the text itself and external sources (de Bruijn et al. 2011; M. Jiang et al. 2011), we used only lemmas, POS tags and chunk tags generated by the modified GENIA tagger. Our approach offers the possibility to construct effective clinical concept annotation systems on a simple feature set, without using dictionaries or ontologies.

Recent studies have shown that the performance of combined or hybrid annotation systems is better than that of any individual systems (Kang et al. 2012). However, our experiment results are not able to support their findings. Only the parallel combination of truecasing and AD outperformed truecasing and AD individually; the performance of the other hybrid NER models fell in between that of the best and worst performing individual methods. This can be attributed to the possibly conflicting contributions of those three pre- or post-processing methods. In our research, different preprocessing or post-processing components are added to the same core information extraction system (i.e., NERsuite), and this could be the reason why the improvements are modest.

The truecasing method improved the concept extraction performance based on both exact and relaxed matching, demonstrating that correct case information is beneficial for entity recognition in clinical records. The expansion and disambiguation of abbreviations were supposed to decrease the size of acronyms and abbreviations in the set of false negatives. Unexpectedly, the false negatives generated by AD are even slightly greater than that from truecasing and the baseline methods (see Table 2.7). The low performance of AD can partly be explained by the fact that the AcroMine Disambiguation tool which we used was trained on MEDLINE abstracts (Okazaki et al. 2010). Training on a suitable abbreviation disambiguation dictionary specifically geared towards terms in clinical records would likely enhance the performance of AcroMine Disambiguation.
Table 2.7: Number of unrecognised abbreviations.  
(\%: Number of unrecognised abbreviations/Total number of abbreviations in the test set*100)

<table>
<thead>
<tr>
<th>Systems</th>
<th>Exact matching</th>
<th></th>
<th></th>
<th></th>
<th>Relaxed matching</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FN</td>
<td>(%)</td>
<td>FP</td>
<td>(%)</td>
<td>FN</td>
<td>(%)</td>
<td>FP</td>
<td>(%)</td>
</tr>
<tr>
<td>Baseline</td>
<td>711</td>
<td>2.14</td>
<td>379</td>
<td>1.14</td>
<td>579</td>
<td>1.75</td>
<td>298</td>
<td>0.90</td>
</tr>
<tr>
<td>T</td>
<td>706</td>
<td>2.13</td>
<td>372</td>
<td>1.12</td>
<td>586</td>
<td>1.77</td>
<td>294</td>
<td>0.89</td>
</tr>
<tr>
<td>AD</td>
<td>724</td>
<td>2.18</td>
<td>472</td>
<td>1.42</td>
<td>565</td>
<td>1.70</td>
<td>372</td>
<td>1.12</td>
</tr>
<tr>
<td>DS</td>
<td>647</td>
<td>1.95</td>
<td>1438</td>
<td>4.34</td>
<td>513</td>
<td>1.55</td>
<td>1270</td>
<td>3.83</td>
</tr>
<tr>
<td>T + AD (S)</td>
<td>707</td>
<td>2.13</td>
<td>477</td>
<td>1.44</td>
<td>558</td>
<td>1.68</td>
<td>377</td>
<td>1.14</td>
</tr>
<tr>
<td>T + AD + DS (S)</td>
<td>650</td>
<td>1.96</td>
<td>1111</td>
<td>3.35</td>
<td>501</td>
<td>1.51</td>
<td>909</td>
<td>2.74</td>
</tr>
<tr>
<td>T + AD (U)</td>
<td>621</td>
<td>1.87</td>
<td>541</td>
<td>1.63</td>
<td>464</td>
<td>1.40</td>
<td>443</td>
<td>1.34</td>
</tr>
<tr>
<td>T + AD + DS (U)</td>
<td>693</td>
<td>2.09</td>
<td>406</td>
<td>1.22</td>
<td>559</td>
<td>1.69</td>
<td>317</td>
<td>0.96</td>
</tr>
</tbody>
</table>

The DS method did not add much value to the performance of the NER models, and in certain cases, even reduced their performance significantly. For example, from Table 2.5 we found an 8.99 and 7.74 percentage points drop in F-score based on exact and relaxed matching, respectively, when we added DS to the sequential combination of truecasing and AD. A possible explanation is that the amount of the training records used to build the distributional thesaurus is too small to cover the full breadth of term occurrence, which limited the thesaurus’ efficacy. Thus, more records need to be involved to build a high-quality distributional thesaurus.

2.5 Summary

This chapter offered a description of the techniques that can be used to improve the performance of the phenotypic concept extraction. A machine learning (ML)-based model for clinical entity recognition was developed and the effects of three pre- and post-processing methods (i.e., truecasing, AD, and DS) individually or being combined using two schemes (i.e., sequential and parallel) were systematically evaluated. Our results indicate that the addition of truecasing achieved the best
performance using exact matching whilst using relaxed matching the maximum F-score was obtained by the parallel combination of the truecasing and AD methods.
Chapter 3

Corpus development

Chronic obstructive pulmonary disease (COPD) is an umbrella term for a range of lung abnormalities characterised by restricted airflow from the lungs, which is only partially reversible (ATS 1995). It has rapidly become one of the major causes of morbidity and mortality worldwide. In 2002, COPD was known to be the fifth-leading cause of death globally, and is predicted to become the third one by 2030 (WHO 2014b).

COPD is a very complex and heterogeneous disease and not every patient responds to all drugs available for treatment (Miravitlles et al. 2013). The identification of subjects who are responsive to therapies for chronic diseases facilitates the provision of the most appropriate treatment and the prevention of unnecessary medications (Miravitlles et al. 2012). Personalised treatment based on phenotypes of each individual patient is ultimately required to ensure that the most suitable therapies are provided (Shen and He 2014). COPD phenotype is defined as “a single or combination of disease attributes that describe differences between individuals with COPD as they relate to meaningful outcomes (symptoms,
exacerbations, response to therapy, rate of disease progression, or death), which allows for well-defined grouping of patients according to their prognostic and therapeutic characteristics (Han et al. 2010). These phenotypes can help clinicians in identifying patients that respond to specific pharmacological interventions (Miravitlles et al. 2013). They are thus the basis of the implementation of personalised treatment, in which different disease characteristics of COPD patients, together with their severity will be the key in choosing the optimal treatment option (Miravitlles et al. 2012). The idea is to recognise that a patient belongs to a particular phenotype group which would enable clinicians to recommend therapy that is known to be suitable to that group.

However, as clinicians at the point of care use free text in describing phenotypes, such information can easily become obscured and inaccessible (Pathak et al. 2013). In order to expedite the process of identifying a given patient’s COPD group, the phenotypic information hidden in these documents needs to be automatically extracted and distilled for the clinicians’ perusal.

Capable of automatically distilling information expressed in natural language within documents, text mining methods can be used to efficiently extract COPD phenotypes of interest. However, no prior work has been reported on developing gold standard annotated corpora for COPD. On the basis of the improved phenotype extraction methods presented in Chapter 2, we develop a text mining-assisted methodology for the gold-standard annotation of COPD phenotypes. With the resulting gold standard corpus, we aim to support the development and evaluation of text mining systems that can ultimately be applied to evidence-based healthcare and clinical decision support systems.

3.1 Literature review

Various corpora have been constructed to support the development of clinical NLP methods. Some contain annotations formed on the basis of document-level tags indicating the specific diseases that clinical reports pertain to. In the 2007 Computational Medicine Challenge data set (Pestian et al. 2007), radiology reports were assigned codes from the ninth revision of the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) terminology (Sickbert-Bennett et al. 2010). In similar corpora, chest X-ray reports were manually labelled
with any of four pneumonia-related concepts (Fiszman et al. 2000) whilst any of 80 possible disease names were assigned to documents in another collection of clinical records (Meystre and Haug 2006) with the assistance of automatic tools MetaMap Transfer (MMTx) (Aronson and Lang 2010a) for concept recognition and NegEx (Chapman et al. 2001) for negation detection. Whilst suitable for evaluating information retrieval methods, such document-level annotations cannot sufficiently support the extraction of phenotypic concepts which are described in clinical records in largely variable ways, making it necessary for automated methods to perform analysis by looking at their actual mentions within text.

Several other clinical corpora were thus enriched with text-bound annotations, which serve as indicators of specific locations of phenotypic concept mentions within text. For instance, all mentions of signs or symptoms, medications and procedures relevant to inflammatory bowel disease were marked up in the corpus developed by South et al. (2009). Specific mentions of diseases and signs or symptoms were similarly annotated under the ShARE scheme (Deléger et al. 2012; Suominen et al. 2013) and additionally linked to terms in the SNOMED-CT. Whilst the scheme developed by Bodenreider (2004) had similar specifications, it is unique in terms of its employment of an automatic tool to accelerate the annotation process. One difficulty encountered by annotators following such a scheme, however, is with manually mapping phenotypic terms to vocabulary concepts, owing to the high degree of variability with which these concepts are expressed in text. For instance, many signs or symptoms (e.g., gradual progressive breathlessness), cannot be fully mapped to any of the existing terms in vocabularies.

Alleviating this issue are schemes which were designed to enrich corpora with finer-grained text-bound annotations. The Clinical e-Science Framework (CLEF) annotation scheme (Roberts et al. 2009), which defined several clinical concept types and relationships, required the decomposition of phrases into their constituent concepts which were then individually assigned concept type labels and linked using any of their defined relationships. Also based on a fine-grained annotation approach is the work by Mungall et al. (2010) on the ontology-driven annotation of interspecies phenotypic information based on the Entity-Quality (EQ) model. Although their work was carried out with the help of the Phenote software (Phenote 2014) for storing, managing and visualising annotations, the entire curation process was done manually, i.e., without the support of any NLP tools. The effort we have undertaken,
in contrast, can be considered as a step towards automating such EQ model-based fine-grained annotation of phenotypic information.

In this regard, our work is unique amongst annotation efforts within the clinical NLP community, but shares similarities with some phenotype curation pipelines employed in the domain of biological systematics. Curators of the Phenoscape project (Dahdul et al. 2010) manually link EQ-encoded phenotypes of fishes to the Zebrafish Model Organism Database using Phenex (Balhoff et al. 2010) which is a tool for managing character-by-taxon matrices, a formal approach used by evolutionary biologists. To accelerate this process, Phenex has been recently enhanced with NLP capabilities (Cui et al. 2012) upon the integration of a text analytic known as CharaParser (Cui 2012). Based on a combination of bootstrapping and syntactic parsing approaches (Cui et al. 2010), CharaParser can automatically annotate structured characteristics of organisms (i.e., phenotypes) in text, but currently does not have full support for linking concepts to ontologies (“Phenoscape - CharaParser” 2014). Also facilitating the semi-automatic curation of systematic literature is GoldenGATE (Sautter et al. 2007), a stand-alone application modelled after the GATE framework (Cunningham et al. 2013), which allows for the combination of various NLP tools into text processing pipelines. It is functionally similar to our Web-based annotation platform Argo in terms of its support for NLP workflow management and manual validation of automatically generated annotations. However, the latter fosters interoperability to a higher degree by conforming to the industry-supported UIMA and allowing workflows to be invoked as Web services (Rak et al. 2014).

By producing our proposed fine-grained phenotype annotations which are linked to ontological concepts, we are representing them in a computable form thus making them suitable for computational applications such as inferencing and semantic search. The Phenomizer tool (Köhler et al. 2009), for instance, has demonstrated the benefits of encoding phenotypic information in a computable format. Leveraging the Human Phenotype Ontology (HPO) (Köhler et al. 2014) whose terms are linked to diseases in the Online Mendelian Inheritance in Man (OMIM) vocabulary (Hamosh et al. 2005), it supports clinicians in making diagnoses by semantically searching for the medical condition that best matches the HPO signs or symptoms given in a query. We envisage that such an application, when integrated with a repository of phenotypes and corresponding clinical
recommendations, e.g., Phenotype Portal (SHARPn 2014) and the Phenotype KnowledgeBase (eMERGE Network 2014), can ultimately assist point-of-care clinicians in more confidently providing personalised treatment to patients. Our work on the annotation of COPD phenotypes aims to support the development of similar applications in the future.

### 3.2 Document collection

Documents from two different domains, i.e., the clinical and the biomedical domains, were collected.

#### 3.2.1 Clinical records

Since clinical records are extremely personal and sensitive, it is always a challenge for scientific researchers to access the corpora of clinical notes. Sharing clinical records with parties outside of hospitals requires a de-identification process, which ensures that all individually identifiable protected health information (PHI) pertaining to names, identification numbers, addresses, phones numbers, and age (specifically those above 90), is removed from the records (Malin 2012; Uzuner et al. 2007). There are several public collections of anonymised clinical records that have become available, such as the MIMIC II Clinical Database.

##### 3.2.1.1 The MIMIC II Clinical Database

Developed to support epidemiologic research and the evaluation of new clinical decision support and monitoring systems, the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) Clinical Database (Goldberger et al. 2000; Saeed et al. 2011) stores comprehensive and detailed clinical information such as discharge summaries, nursing progress notes, radiology reports, microbiology test results and comorbidity scores. Its most current version contains the above types of information for 36,095 hospital admissions of 32,536 adult intensive care unit (ICU) patients. In building our corpus, we exploited this publicly accessible database and selected 1,000 COPD-relevant clinical records based on criteria suggested by two experts on COPD.
3.2.1.2 Selection criteria

Based on the recommendation of our domain experts, we looked into the available details relating to COPD comorbidities and the sputum specimens used in microbiological tests.

For the former, we queried the database for patient admissions during which COPD and any other 29 comorbidities were recorded in the hospital. Number of records of patients with each ailment and its relationship with COPD are shown in Table 3.1 and the distribution is also visualised in Figure 3.1. The five most frequently co-occurring diseases with COPD are determined. These are, namely, pulmonary circulatory disorders, obesity, congestive heart failure, peripheral vascular disorders and cardiac arrhythmias.

![Figure 3.1: Distribution of records of each disease without/with COPD.](image)
Table 3.1: Number of MIMIC clinical records of each disease and its relationship with COPD.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of records</th>
<th>Number of records with COPD</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPD</td>
<td>3924</td>
<td>3924</td>
<td>1</td>
</tr>
<tr>
<td>Pulmonary circulatory disorders</td>
<td>505</td>
<td>173</td>
<td>0.3426</td>
</tr>
<tr>
<td>Obesity</td>
<td>384</td>
<td>106</td>
<td>0.2760</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>4616</td>
<td>1178</td>
<td>0.2552</td>
</tr>
<tr>
<td>Peripheral vascular disorders</td>
<td>1862</td>
<td>433</td>
<td>0.2325</td>
</tr>
<tr>
<td>Cardiac arrhythmias</td>
<td>4555</td>
<td>957</td>
<td>0.2101</td>
</tr>
<tr>
<td>Depression</td>
<td>1173</td>
<td>241</td>
<td>0.2055</td>
</tr>
<tr>
<td>Solid tumour without metastasis</td>
<td>3585</td>
<td>728</td>
<td>0.2031</td>
</tr>
<tr>
<td>Valvular heart disease</td>
<td>1974</td>
<td>400</td>
<td>0.2026</td>
</tr>
<tr>
<td>Phychoses</td>
<td>814</td>
<td>160</td>
<td>0.1966</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>500</td>
<td>97</td>
<td>0.1940</td>
</tr>
<tr>
<td>Uncomplicated diabetes</td>
<td>4477</td>
<td>825</td>
<td>0.1843</td>
</tr>
<tr>
<td>Fluid and electrolyte disorders</td>
<td>5670</td>
<td>1040</td>
<td>0.1834</td>
</tr>
<tr>
<td>Deficiency anemias</td>
<td>2593</td>
<td>473</td>
<td>0.1824</td>
</tr>
<tr>
<td>Hypertension</td>
<td>7820</td>
<td>1405</td>
<td>0.1797</td>
</tr>
</tbody>
</table>
Number of MIMIC clinical records of each disease and its relationship with COPD. (Continued from previous page)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Number of Records</th>
<th>Number of COPD</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS</td>
<td>151</td>
<td>26</td>
<td>0.1722</td>
</tr>
<tr>
<td>Weight loss</td>
<td>564</td>
<td>95</td>
<td>0.1684</td>
</tr>
<tr>
<td>Peptic ulcer disease</td>
<td>132</td>
<td>22</td>
<td>0.1667</td>
</tr>
<tr>
<td>Renal Failure</td>
<td>1237</td>
<td>203</td>
<td>0.1641</td>
</tr>
<tr>
<td>Metastatic cancer</td>
<td>1093</td>
<td>178</td>
<td>0.1629</td>
</tr>
<tr>
<td>Complicated diabetes</td>
<td>1112</td>
<td>173</td>
<td>0.1556</td>
</tr>
<tr>
<td>Other neurological disorders</td>
<td>807</td>
<td>125</td>
<td>0.1549</td>
</tr>
<tr>
<td>Paralysis</td>
<td>324</td>
<td>49</td>
<td>0.1512</td>
</tr>
<tr>
<td>Coagulopathy</td>
<td>1434</td>
<td>214</td>
<td>0.1492</td>
</tr>
<tr>
<td>Liver disease</td>
<td>1139</td>
<td>167</td>
<td>0.1466</td>
</tr>
<tr>
<td>Alcohol abuse</td>
<td>1379</td>
<td>182</td>
<td>0.1320</td>
</tr>
<tr>
<td>Drug abuse</td>
<td>733</td>
<td>94</td>
<td>0.1282</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>383</td>
<td>45</td>
<td>0.11753</td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>4905</td>
<td>367</td>
<td>0.0748</td>
</tr>
<tr>
<td>Blood loss anemia</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Charlson comorbidity index

The Charlson comorbidity index is a method of categorising comorbidities of patients based on the ICD diagnosis codes found in administrative data. Each comorbidity category has an associated weight (1, 2, 3, or 6) depending on the risk of mortality, and the sum of all the weights results in a single comorbidity score. A score of zero indicates that no comorbidities were found. The higher the score, the more likely the predicted outcome will result in mortality. 30 diseases stated in MIMIC II data set and associated weight are as follows:

- 1: COPD, pulmonary circulatory disorders, obesity, congestive heart failure, peripheral vascular disorders, cardiac arrhythmias, depression, valvular heart disease, phychoses, rheumatoid arthritis, uncomplicated diabetes, fluid and electrolyte disorders, deficiency anemias, hypertension, weight loss, peptic ulcer disease, complicated diabetes, other neurological disorders, paralysis, coagulopathy, chronic liver disease, alcohol abuse, drug abuse, hypothyroidism, or blood loss anemia;

- 2: Solid tumour without metastasis, renal failure, or lymphoma;

- 3: Moderate or severe liver disease;

- 6: Acquired immune deficiency syndrome (AIDS), or metastatic cancer.
Charlson comorbidity index of those 30 clinical conditions is also calculated (as shown in Figure 3.2), and we found that the patients with COPD tend to suffer from more than one comorbidity and thus have a higher risk of mortality.

![Charlson comorbidity index of comorbidities with/without COPD](image)

Figure 3.2: Charlson comorbidity index of comorbidities with/without COPD.

For the sputum specimens used in microbiological tests, we examined the database entries for associations between bacteria and four dimensions: the patients’ length of stay in hospital, length of stay in the ICU, in-hospital mortality and the five comorbidities we have previously shortlisted. We identified six types of bacteria which are most frequently associated with longer hospital/ICU stays, mortality, and comorbidities, i.e., coagulase-positive *S. aureus*, gram-negative rods, *P. aeruginosa*, *K. pneumoniae*, *E. coli* and *A. fumigates*.

We restricted the pool of documents for consideration to those meeting these two criteria, i.e., clinical records of COPD patients having any of the five comorbidities and whose sputum specimens have any of the six bacteria. Out of an initial set of 22,265 matching records, we randomly selected 1,000 for our corpus. The resulting document collection (as shown in Table 3.2) consists of six discharge summaries, seven medical doctor’s notes, 226 radiology reports and 761 nursing progress notes representing a total of 296 patients.
Table 3.2: Distribution of 1000 MIMIC II clinical records over different document type.

<table>
<thead>
<tr>
<th>Document type</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discharge summary</td>
<td>6</td>
</tr>
<tr>
<td>Medical doctor’s note</td>
<td>7</td>
</tr>
<tr>
<td>Radiology report</td>
<td>226</td>
</tr>
<tr>
<td>Nursing progress note</td>
<td>761</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1000 (representing 296 patients)</strong></td>
</tr>
</tbody>
</table>

### 3.2.2.3 COPD patient corpus

There are obvious differences in how the American and British hospitals work (Senior and Anthonisen 1998). The British National Health Service (NHS) standards are set forth by the National Institute for Health and Care Excellence (NICE) that establish guidelines and standards for a wide range of diseases. For example, asthma patients who are seen at emergency or after hours care must be seen within the next two days by their primary healthcare provider. To compare, the United States (US) does not utilize care guidelines in the same manner that the NHS does. The clinical guidelines are set forth by the Agency for Healthcare Research and Quality (AHRQ) who define standards for areas like diagnostic pharmaceuticals, testing and treatment of ailments.

We want to develop methods that can eventually be of use to clinicians in the United Kingdom (UK). However, the MIMIC II records that have been used in our research because of their rich annotations were collected from the hospitals in the US. Clinical records of British patients, therefore, need to be collected.

Moreover, the public could not straightforwardly access the documents in MIMIC II as one has to complete a three-hour online course about protecting human research participants and fill in a restricted data use agreement before getting the permission (MIMIC II 2012).

Therefore, with the cooperation with our clinical experts, we successfully collected the clinical records of 50 patients with COPD in the UK who were born from 1927 to 1985, among which there are 22 females and 28 males. The corpus consists of four types, i.e., diagnosis, procedures, letters and results, which can be
related through the Patient ID number. The description of each category is shown in Table 3.3.

### 3.2.2 Biomedical literature

In forming our biomedical corpus, we collected pertinent journal articles from the PubMed Central (PMC) Open Access subset. As a preliminary step, we retrieved a list of journals which are most relevant to COPD by querying journal titles in PMC Open Access using the keywords “chronic”, “obstructive”, “pulmonary”, “disease”, “respiratory” and “lung” (shown in Table 3.4). This resulted in 10 journal titles whose archives were then searched for the keywords “chronic obstructive pulmonary disease” and “COPD”. A total of 974 full-text articles were retrieved in this manner. The journal titles and article distribution over them are shown in Figure 3.3

![Figure 3.3: Distribution of COPD-relevant articles over COPD-focused journals.](image-url)
Table 3.3: Description of collected clinical records of COPD patients.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of records</th>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>253</td>
<td>1) PatNo</td>
<td>Unique patient number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Description</td>
<td>Diagnosis or diseases of patients, e.g., seropositive rheumatoid arthritis, lung cancer, and COPD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Code</td>
<td>ICD-9-CM code of diagnosis</td>
</tr>
<tr>
<td>Procedures</td>
<td>46</td>
<td>1) PatNo</td>
<td>Same as in Diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Description</td>
<td>Treatments carried out for the patients, e.g., coronary angiogram, bariatric surgery, and pacemaker implant dual chamber</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Code</td>
<td>ICD-9-CM code of the procedures</td>
</tr>
<tr>
<td>Letters</td>
<td>48,697</td>
<td>1) PatNo</td>
<td>Same as in Diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) DocumentName</td>
<td>Records type, e.g., discharge summary, clinical note, nursing evaluation, clinical observations, senior review, and risk assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Detailtext</td>
<td>Detailed description</td>
</tr>
<tr>
<td>Results</td>
<td>82,499</td>
<td>1) PatNo</td>
<td>Same as in Diagnosis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) ItemName</td>
<td>Biomedical parameter name, e.g., protein, sodium, phosphate, hematocrit (HCT), and white blood cell (WBC)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Value</td>
<td>Parameter value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4) UnitofMeasure</td>
<td>Measurement unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5) ResText</td>
<td>Notes and diagnosis made for each specific biomedical parameter</td>
</tr>
</tbody>
</table>
Table 3.4: PMC journals relevant to COPD.

<table>
<thead>
<tr>
<th>ISSN</th>
<th>Title</th>
<th>Volumes in PMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1471-2466</td>
<td>BMC Pulmonary Medicine</td>
<td>v.14 2014</td>
</tr>
<tr>
<td>1541-7891</td>
<td>Cardiopulmonary Physical Therapy Journal</td>
<td>v.24(3) Sep 2013 v.19 2008</td>
</tr>
<tr>
<td>1179-5484</td>
<td>Clinical Medicine Insights. Circulatory, Respiratory and Pulmonary Medicine (v.2;2008)</td>
<td>v.8 2014 v.4 2010</td>
</tr>
<tr>
<td>1176-9106</td>
<td>International Journal of Chronic Obstructive Pulmonary Disease</td>
<td>v.9 2014 v.1 2006</td>
</tr>
<tr>
<td>2045-8932</td>
<td>Pulmonary Circulation</td>
<td>v.3(2) Apr-Jun 2013 v.1 2011</td>
</tr>
<tr>
<td>2282-8419</td>
<td>Heart, Lung and Vessels</td>
<td>v.5(4) 2013 v.5 2013</td>
</tr>
<tr>
<td>1545-1151</td>
<td>Preventing Chronic Disease</td>
<td>v.11 2014 v.1 2014</td>
</tr>
<tr>
<td>2040-6223</td>
<td>Therapeutic Advances in Chronic Disease</td>
<td>v.5(1) Mar 2014 v.1 2010</td>
</tr>
</tbody>
</table>
Upon consideration of our constraints in terms of resources such as time and personnel, we decided to trim down the document set to 30 full articles. This was carried out by compiling a list of COPD phenotypes based on the combination of terms given by our domain experts and those automatically extracted by Termine (Frantzi et al. 2000) from the COPD guidelines published jointly by the American Thoracic Society and the European Respiratory Society in 2004 (ATS and ERS 2004). The resulting term list (provided as Appendix A) contains 1,925 COPD phenotypes which were matched against the content of the initial set of 974 articles, and the query used for searching is displayed below.

- **Query building**
  ```
  cat copd_keywords.txt | perl -pe 's/\n/" OR "/g;

db = pmc (Database: PMC, default=pubmed)

  ```

In order to ensure that the documents in our corpus are representative of the widest possible range of COPD phenotypes, we ranked the documents according to decreasing number of their contained unique matches. We then selected the 30 top-ranked articles as the final document set for our corpus, as shown in Table 3.5.
Table 3.5: 30 top-ranked biomedical articles.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Counts of unique phenotypes</th>
<th>Counts of all phenotypes</th>
<th>File title in PMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>168</td>
<td>630</td>
<td>copd-2-493</td>
</tr>
<tr>
<td>2</td>
<td>137</td>
<td>643</td>
<td>copd-6-551</td>
</tr>
<tr>
<td>3</td>
<td>136</td>
<td>692</td>
<td>copd-1-381</td>
</tr>
<tr>
<td>4</td>
<td>135</td>
<td>629</td>
<td>PM2011-257496</td>
</tr>
<tr>
<td>5</td>
<td>132</td>
<td>600</td>
<td>copd-6-199</td>
</tr>
<tr>
<td>6</td>
<td>128</td>
<td>727</td>
<td>copd-9-349</td>
</tr>
<tr>
<td>7</td>
<td>126</td>
<td>787</td>
<td>copd-4-321</td>
</tr>
<tr>
<td>8</td>
<td>118</td>
<td>943</td>
<td>copd-4-203</td>
</tr>
<tr>
<td>9</td>
<td>114</td>
<td>633</td>
<td>COPD-3-563</td>
</tr>
<tr>
<td>10</td>
<td>112</td>
<td>649</td>
<td>copd-1-219</td>
</tr>
<tr>
<td>11</td>
<td>112</td>
<td>602</td>
<td>copd-4-351</td>
</tr>
<tr>
<td>12</td>
<td>112</td>
<td>631</td>
<td>PC-3-5</td>
</tr>
<tr>
<td>13</td>
<td>111</td>
<td>449</td>
<td>COPD-3-671</td>
</tr>
<tr>
<td>14</td>
<td>110</td>
<td>739</td>
<td>copd-0301-55</td>
</tr>
<tr>
<td>15</td>
<td>110</td>
<td>737</td>
<td>copd-9-187</td>
</tr>
<tr>
<td>16</td>
<td>108</td>
<td>559</td>
<td>COPD-3-605</td>
</tr>
</tbody>
</table>
30 top-ranked biomedical articles. (Continued from previous page)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>107</td>
<td>617</td>
<td>copd-1-409</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>107</td>
<td>420</td>
<td>copd-9-027</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>106</td>
<td>487</td>
<td>copd-6-047</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>106</td>
<td>639</td>
<td>PM2012-203952</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>103</td>
<td>579</td>
<td>copd-5-153</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>102</td>
<td>900</td>
<td>COPD-3-585</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>102</td>
<td>477</td>
<td>copd-9-139</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>101</td>
<td>316</td>
<td>1471-2466-6-27</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>99</td>
<td>683</td>
<td>copd-1-3</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>99</td>
<td>765</td>
<td>copd-2-205</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>96</td>
<td>645</td>
<td>ccrpm-5-2011-057</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>95</td>
<td>511</td>
<td>ccrpm-7-2013-017</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>95</td>
<td>574</td>
<td>copd-5-165</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>95</td>
<td>1110</td>
<td>PC-3-179</td>
<td></td>
</tr>
</tbody>
</table>
3.3 Annotation scheme and procedure

3.3.1 Annotation scheme

To capture and represent phenotypic information, we developed a typology of clinical concepts (Table 3.6) taking inspiration from the definition of COPD phenotypes previously proposed (Han et al. 2010). After reviewing the semantic representations used in previous clinical annotation efforts, we decided to adapt and harmonise concept types from the annotation schemes applied to the 2010 i2b2/VA Shared Task data set (Uzuner et al. 2011) and the PhenoCHF corpus (Alnazzawi et al. 2014). In the former, as we mentioned before, clinical concepts of interest are organised into three broad categories, i.e., Problem, Treatment and Test. However, according to the discussions with our domain experts, it is necessary to develop a finer-grained typology for better COPD phenotype extraction. Therefore, we adapted and organised some of the semantic types used in the annotation of congestive heart failure phenotypes in the PhenoCHF corpus, shown in the Table 3.6 with asterisks, under the upper-level types of the i2b2/VA scheme. More examples can be found in Appendix B.

From the examples in Table 3.6, we can notice that many phenotypes span full phrases, such as *increased levels of the C-reactive protein*, and *decrease in rate of lung function*. In order to produce computable and expressive annotations, previous researchers proposed some schemes using highly structured representations. For example, in the CLEF (Roberts et al. 2009), the phrase *decrease in rate of lung function* will be broken down into two segments, *decrease in rate* and *lung function*; *decrease in rate* will be labelled as a condition, *lung function* as locus and then they will be linked via a “has_location” relation. The EQ model, similarly, will decompose the phrase into *lung function* (E) and *decrease (Q) in rate* in our example respectively. Tree structures were used to present the extended named entities in the work of Dinarelli and Rosset (2012). It is ideal to use such representations for automatic knowledge inference, representation and detection. However, domain experts are quite likely to find difficulty in capturing those annotations and relationships as they may lack the required background knowledge in linguistics.
Table 3.6: Annotation scheme for capturing COPD phenotypes.
(*: Semantic types adapted from the PhenoCHF scheme)

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Problem</td>
<td>an overall category for any COPD indications of concern</td>
<td>frequent exacerbator</td>
</tr>
<tr>
<td>a) Condition*</td>
<td>any disease or medical condition; includes COPD comorbidities</td>
<td>emphysema, pulmonary vascular disease, asthma, congestive heart failure</td>
</tr>
<tr>
<td>b) RiskFactor*</td>
<td>a phenotype signifying a patient’s increased chances of having COPD</td>
<td>increased levels of the C-reactive protein, alpha1 antitrypsin deficiency</td>
</tr>
<tr>
<td>i) SignOrSymptom*</td>
<td>an observable irregularity manifested by a COPD patient</td>
<td>chronic cough, shortness of breath, purulent sputum production</td>
</tr>
<tr>
<td>ii) IndividualBehaviour*</td>
<td>a patient’s habits leading to susceptibility of having COPD</td>
<td>smoking for 25 years</td>
</tr>
<tr>
<td>iii) TestResult*</td>
<td>findings based on COPD-relevant examinations</td>
<td>decrease in rate of lung function, FEV1 45% predicted</td>
</tr>
<tr>
<td>2) Treatment</td>
<td>any medication, therapy or program for treating COPD</td>
<td>oxygen therapy, pulmonary rehabilitation, pursed lips breathing</td>
</tr>
<tr>
<td>3) Test</td>
<td>an overall category for any COPD-relevant examinations or measures/parameters</td>
<td>increased compliance of the lung, FEV1, FEV1/FVC ratio</td>
</tr>
<tr>
<td>a) RadiologicalTest</td>
<td>any of the radiological tests for detecting COPD</td>
<td>computed tomography scanning, high resolution computed tomography</td>
</tr>
<tr>
<td>b) MicrobiologicalTest</td>
<td>an examination of a COPD-relevant specimen</td>
<td>complete blood count</td>
</tr>
<tr>
<td>c) PhysiologicalTest</td>
<td>a measurement of a COPD patient’s capacity to exercise</td>
<td>6-min walking distance</td>
</tr>
</tbody>
</table>
We therefore introduced an annotation methodology that is simple for human annotators, yet produces structured, computable annotations. Some examples are summarised in Table 3.7. The simplicity lies in the fact that the annotators are only asked to simply mark up text spans, i.e., the COPD phenotypes. However, we can still produce structured annotations with the application of various text mining tools which are capable of recognising the concepts underlying the COPD phenotypes. For example, the underlying concepts decrease in rate, lung, and function of COPD phenotype decrease in rate of lung function can be annotated using our methodology. Furthermore, these underlying concepts can also be mapped to pertinent ontologies, such as the Phenotype and Trait Ontology (PATO) (OBO 2014) and Uber Anatomy Ontology (UBERON) (Mungall et al. 2012), to make them computable, e.g., decreased rate (PATO: 0000911), lung (UBERON: 0002048), and function (PATO: 0000173).

Table 3.7: Examples of phenotypes represented using our annotation scheme.

<table>
<thead>
<tr>
<th>COPD phenotype</th>
<th>Underlying concept</th>
<th>Linked ontological concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>decrease in rate of lung function</td>
<td>decrease in rate lung function</td>
<td>decreased rate (PATO:0000911) lung (UBERON:0002048) function (PATO:0000173)</td>
</tr>
<tr>
<td>chronic airways obstruction</td>
<td>chronic airways obstruction</td>
<td>chronic (PATO:0001863) respiratory airway (UBERON:0001005) obstructed (PATO:0000648)</td>
</tr>
<tr>
<td>chronic bronchitis</td>
<td>N/A</td>
<td>chronic bronchitis (DOID:6132)</td>
</tr>
</tbody>
</table>

3.3.2 Text mining-assisted annotation

The contribution of text mining tools to document annotation is two-fold. Firstly, text mining will accelerate the manual annotation process, and secondly, the tools can provide the granular concept annotations we are aiming for to make the annotations in our corpus more expressive.
We used Argo (See section 2.3.2) that is an interoperable Web-based text mining platform to both integrate our elementary analytics into a processing workflow and to manage its execution. From Argo’s rich library, we selected the components (as described in Table 3.8) which are most suitable for our task’s requirements and arrange them in a semi-automatic annotation workflow (as depicted in Figure 3.4).
Table 3.8: Argo components used in our annotation workflow and their descriptions.

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Input annotation(s)</th>
<th>Output annotation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Reader</td>
<td>Reads documents from the Argo folders</td>
<td>None declared</td>
<td>None declared</td>
</tr>
<tr>
<td>Cafetiere Sentence Splitter</td>
<td>Uses a set of heuristics and patterns to find sentence boundaries</td>
<td>Text</td>
<td>Sentence</td>
</tr>
<tr>
<td>GENIA Tagger</td>
<td>Tags biological named entities, i.e., proteins, cell lines, cell types, DNAs, and RNAs. It contains tokeniser, part-of-speech tagger, and shallow parser. The models were trained on the GENIA corpus.</td>
<td>Sentence, Token</td>
<td>RichToken, Chunk, RNA, Protein, DNA, CellType, CellLine</td>
</tr>
<tr>
<td>Brown Dictionary Feature Extractor</td>
<td>Extracts dictionary features based on a dictionary of Brown Clusters created from MEDLINE articles</td>
<td>Sentence, Chunk, AlternateToken</td>
<td>DictionaryFeatures</td>
</tr>
<tr>
<td>OBO Anatomy Dictionary Feature Extractor</td>
<td>Extracts dictionary features based on a dictionary of Anatomical Entities extracted from OBO files found in the OBOFoundry</td>
<td>Sentence, Chunk, AlternateToken</td>
<td>DictionaryFeatures</td>
</tr>
<tr>
<td>UMLS Dictionary Feature Extractor</td>
<td>Extracts dictionary features from a dictionary created from the UMLS Metathesaurus</td>
<td>Sentence, Chunk, AlternateToken</td>
<td>DictionaryFeatures</td>
</tr>
</tbody>
</table>
Argo components used in our annotation workflow and their descriptions. (Continued from previous page)

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Mentioned Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NERsuite Tagger</strong></td>
<td>The modification of the original NERsuite tagger by adding Truecase Asciiifier (as described in Chapter 2). Uses dictionary features.</td>
<td>Chemical, Gene, Protein, Disease, Clinical Concept, Enzyme, Mechanism, PKParameter, PKValue, Change, NegationCue, Person, GeoPoliticalEntity, Organisation, Location, Facility, Vehicle, Weapon, ValueMention, Timex2Mention</td>
</tr>
<tr>
<td><strong>Chemical Entity Recogniser</strong></td>
<td>A named entity recogniser capable of annotating names of chemicals, drugs and metabolites. Built on top of the NERsuite package</td>
<td>Chemical, Drug, Metabolite</td>
</tr>
<tr>
<td><strong>Anatomical Entity Tagger</strong></td>
<td>Tags anatomical entities using Brown, UMLS and OBO Anatomy dictionary features</td>
<td>NamedEntity</td>
</tr>
</tbody>
</table>
Argo components used in our annotation workflow and their descriptions. (Continued from previous page)

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>SourceDocumentInformation</th>
<th>None declared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual Annotation Editor</td>
<td>A visual editor of feature structures. It allows for viewing as well as editing (i.e., creating, deleting, updating) all feature structures. Text-span annotations can also be marked directly in text.</td>
<td>SourceDocumentInformation</td>
<td>None declared</td>
</tr>
<tr>
<td>XMI Writer</td>
<td>Serialises entire common annotation structures (CAS) to XMI format.</td>
<td>SourceDocumentInformation</td>
<td>None declared</td>
</tr>
</tbody>
</table>
Figure 3.4: Our semi-automatic annotation workflow established in Argo.
Clinical or biomedical documents are first read in by a Document Reader, and transformed to sentence-splitted documents through Cafetiere Sentence Splitter. GENIA Tagger is then applied to produce the POS and chunk tags, lemmas, as well as the mentions of proteins (Tsuruoka et al. 2005). After running the syntactic tools, four branches of concept recognisers are employed to generate different annotations.

According to our experiment results for phenotype extraction described in Chapter 2, the best performance for exact matching was achieved by adding truecasing to NERsuite. Therefore, in this workflow, we developed a new NERsuite Tagger which is the combination of the original NERsuite and Truecase Ascifler components.

The NER tagger labelled NERsuite Tagger (2) is trained on the 2010 i2b2/VA challenge training set (Fu and Ananiadou 2014). This NER is employed to provide domain experts with automatically generated cues (i.e., Problem, Treatment, and TestOrMeasure) which could aid them in marking up full phrases describing COPD phenotypes. The NER tagger labelled NERsuite Tagger is applied to identify disease mentions using a model trained on the NCBI Disease corpus (Dögan et al. 2014). Drug name recognition is performed by the Chemical Entity Recogniser (Batista-Navarro et al. 2013) trained on the Drug-Drug Interaction (DDI) corpus (Herrero-Zazo et al. 2013) in the third branch. Fourthly, Anatomical Entity Tagger based on the features extracted by Brown, OBO Anatomy and UMLS Dictionary Feature Extractors is used to recognise anatomical concepts (Pyysalo and Ananiadou 2013). The annotations generated by these four kinds of concept recognisers are then collected and combined. Domain experts can easily modify such automatically generated annotations through the graphical user interface (GUI) of the Manual Annotation Editor. Figure 3.5 shows screenshots of two annotated articles, i.e., biomedical and clinical documents. In biomedical literature, for example, *peripheral muscle dysfunction* is recognised as Problem; meanwhile *muscle dysfunction* and *muscle* are granularly identified as Disease and Anatomical Entity, respectively. In clinical text, *ACE inhibitor* is recognised as Treatment as well as Drug. The annotated articles are finally stored in the Extensible Markup Language (XML) Metadata Interchange standard format via the XML Metadata Interchange (XMI) Writer component.
Together with peripheral muscle dysfunction, the major factor that limit exercise tolerance in these patients is the development of DH and the concurrent mechanical constraints on ventilation that contribute importantly to perceived respiratory discomfort (Diaz et al. 2000; Puente-Maestu et al. 2005). Dynamic hyperinflation is evident even with short exercise bouts in interval training (Vogiatzis et al. 2004). In one study, a correlation was found between resting hyperinflation (measured as RV/TLC%) and the increase in endurance time after 8 weeks of leg muscle rehabilitation (Puente-Maestu et al. 2003). In another study, IC was found as a significant predictor of the long term effects after a rehabilitation program, but in a multiple logistic regression model, only pressure of carbon dioxide (PaCO2) was identified as predictor for the maintenance of improvement in health-related quality of life over one year (Nishiyama et al. 2005). Certain variables closely linked to DH (ODonnell et al. 1993, 2001, 2002; Diaz et al. 2000, Marin et al. 2001) such as exercise capacity (maximal oxygen uptake) (Gerardi et al. 1996; Bowen et al. 2000, Myers et al. 2002; Hiraga et al. 2003; Ogas et al. 2003) or 6 minute walking distance (Hinto-Piata et al. 2004), dyspnea (Cetti et al. 2004), and oxygen desaturation (Nishimura et al. 2002; Hiraga et al. 2003; Tojo et al. 2005) during exercise have been shown to be powerful predictors of survival in COPD patients (Table 1). A recent study of 689 COPD patients, with a mean follow up of 34 month, showed that the IC/TLC was
Figure 3.5: GUI of the Manual Annotation Editor component

(a) Segment of one annotated biomedical article;
(b) Annotations of *peripheral muscle dysfunction* (as Problem), *muscle dysfunction* (as Disease) and *muscle* (as NamedEntity);
(c) Segment of one annotated clinical record;
(d) Annotations of *ACE inhibitor* (as Treatment and Drug).
3.4 Results and discussion

3.4.1 Evaluation in the clinical domain

Numbers of occurrences computed over the annotations generated for 1,000 documents of the corpus by the automatic components in our workflow (i.e., the numbers of occurrences of instances of phenotypic expression types (i.e., Problem, Treatment and Test), of elementary/ granular concept types (i.e., anatomical entity, disease, protein and drug), and the overlaps between them) are presented in Table 3.9. The number of overlaps signifies that a considerable percentage (40%) of the phenotypic expressions can be decomposed into granular concepts. However, it also shows that the majority do not contain any instances of our shortlisted elementary types. These numbers provide us with an insight into the volume of annotated phenotypes that could potentially result from our curation effort and the corpus we are building will be a valuable knowledge-rich resource for mining COPD phenotypes from text. It may, for example, be utilized in the development of text mining tools which can automatically demarcate text spans pertaining to COPD phenotypes. Furthermore, our corpus may enable the building of tools that will link such text spans to ontological concepts in order to integrate phenotypic expressions with their semantics.

Table 3.9: Numbers of occurrences of automatically generated annotations in the clinical corpus. (Values in brackets indicate counts of unique instances)

<table>
<thead>
<tr>
<th>Phenotypic expression type</th>
<th>Elementary concept type</th>
<th>Number of instances of phenotypic expression type</th>
<th>Number of instances of elementary concept type</th>
<th>Number of overlaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>anatomical entity</td>
<td>7,254 (4,723)</td>
<td>4,979 (819)</td>
<td>1,841 (533)</td>
</tr>
<tr>
<td>Problem</td>
<td>disease</td>
<td>7,254 (4,723)</td>
<td>3,114 (839)</td>
<td>1,730 (581)</td>
</tr>
<tr>
<td>Problem</td>
<td>protein</td>
<td>7,254 (4,723)</td>
<td>3,861 (1,694)</td>
<td>386 (122)</td>
</tr>
<tr>
<td>Treatment</td>
<td>drug</td>
<td>5,099 (2,548)</td>
<td>2,036 (540)</td>
<td>944 (208)</td>
</tr>
<tr>
<td>Test</td>
<td>N/A</td>
<td>3,726 (1,647)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
3.4.2 Evaluation in the biomedical domain

The evaluation on our corpus of clinical records from the MIMIC II Clinical Database, because of their rich annotations, has been described in the previous section. However, it still does not mean that the public can straightforwardly access those documents. For access to the content of the MIMIC II Clinical Database, one has to complete a three-hour online course about protecting human research participants and fill in a restricted data use agreement before getting the permission (MIMIC II 2012).

Moreover, our UK-based expert collaborators (i.e., stakeholders who will incorporate our text mining technology into their systems in the near future) also pointed out that there are substantial discrepancies between the hospital system in the US (on which MIMIC II is focussed) and that in the UK. After considering their advice, we decided to utilise scientific articles from various COPD-relevant journals, rather than build a corpus of clinical records which are highly US-specific. As previous work demonstrated techniques which successfully extracted information from unseen data even if the training/development data used was of a different document type (Xu et al. 2010), a gold standard corpus of full scientific articles should still allow for the development of phenotype extraction tools for clinical records.

After applying the Argo workflow described in Figure 3.6 on our biomedical articles, we asked one of our collaborating domain experts to manually validate the automatically generated annotations. In this section, we present the results of two types of evaluation. Firstly, the quality of the Argo-generated concept annotations was measured by comparing them against gold standard data, i.e., the annotations manually validated by the domain expert. Secondly, we carried out an evaluation of the gold standard annotations that we have obtained thus far by utilising them in the development of machine learning-based concept recognisers. Table 3.10 presents the number of unique concepts for each type, as manually annotated by our domain expert. One can see that the most prevalent types are Treatment, RiskFactor, MedicalCondition, TestOrMeasure, Drug and AnatomicalConcept (in order of decreasing frequency).
Table 3.10: Number of unique concepts for each type in the biomedical corpus.

<table>
<thead>
<tr>
<th>Concept type</th>
<th>Number of unique concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>1,430</td>
</tr>
<tr>
<td>RiskFactor</td>
<td>1,345</td>
</tr>
<tr>
<td>MedicalCondition</td>
<td>1,279</td>
</tr>
<tr>
<td>TestOrMeasure</td>
<td>783</td>
</tr>
<tr>
<td>Drug</td>
<td>642</td>
</tr>
<tr>
<td>AnatomicalConcept</td>
<td>336</td>
</tr>
<tr>
<td>Quality</td>
<td>182</td>
</tr>
<tr>
<td>Protein</td>
<td>129</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6,216</strong></td>
</tr>
</tbody>
</table>

Table 3.11 depicts the evaluation of Argo’s automatically generated annotations against the gold standard, presented by concept type. Micro and macro-average measurements were used to average the results of each concept type to give a view on the general performance.

- **Micro-average measurement**

Micro-average of precision, recall and F-score are computed by the average of the sum of the individual true positives, false positives, and false negatives using different data sets, as shown below in Equation 3.1, 3.2 and 3.3, where $i$ denotes the different set of concept types and $n$ is the total number of the data set.

\[
Precision_{\text{micro-average}} (P_{\text{micro-average}}) = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \quad (3.1)
\]

\[
Recall_{\text{micro-average}} (R_{\text{micro-average}}) = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \quad (3.2)
\]
\[
F - \text{score}_{\text{micro-average}} = \frac{2 \times P_{\text{micro-average}} \times R_{\text{micro-average}}}{P_{\text{micro-average}} + R_{\text{micro-average}}} 
\]  

(3.3)

- **Macro-average measurement**

Macro-average result is the straightforward average of the precision and recall of the system on different concept sets as explained in Equation 3.4, 3.5 and 3.6.

\[
\text{Precision}_{\text{macro-average}}(P_{\text{macro-average}}) = \frac{1}{n} \sum_{i=1}^{n} \text{Precision}_i 
\]  

(3.4)

\[
\text{Recall}_{\text{macro-average}}(R_{\text{macro-average}}) = \frac{1}{n} \sum_{i=1}^{n} \text{Recall}_i 
\]  

(3.5)

\[
F - \text{score}_{\text{macro-average}} = \frac{2 \times P_{\text{macro-average}} \times R_{\text{macro-average}}}{P_{\text{macro-average}} + R_{\text{macro-average}}} 
\]  

(3.6)

Only the five most frequently occurring concept types (which are common between the manually validated annotations we have at hand and the automatically generated annotations) were included in the evaluation. Exact and relaxed matching were both evaluated. We note that for a given phenotypic expression, not only the full string is being evaluated, but also each of its subsumed concepts. It can be observed from Table 3.11 that in general, the semi-automatic workflow obtains unsatisfactory performance using exact matching. After performing some error analysis, we observed that the majority of discrepancies were brought about by the incorrect inclusion or exclusion of articles or modifiers in noun phrases, e.g., \textit{phosphodiesterase inhibitor} (for \textit{a nonselective phosphodiesterase inhibitor}), \textit{an acute exacerbation} (for \textit{acute exacerbation}). Thus we next employed relaxed matching, which revealed that the semi-automatic workflow obtains moderate performance over all evaluated concept types (except for TestOrMeasure). The comparison of F-score measurements of exact and relaxed matching is shown in Figure 3.6.
Table 3.11: Evaluation of annotations automatically generated by the text mining-assisted workflow against gold standard data.

<table>
<thead>
<tr>
<th>Concept type</th>
<th>Exact matching</th>
<th></th>
<th>Relaxed matching</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
<td>Precision</td>
</tr>
<tr>
<td>AnatomicalConcept</td>
<td>0.1923</td>
<td>0.7527</td>
<td>0.3063</td>
<td>0.2814</td>
</tr>
<tr>
<td>Drug</td>
<td>0.5861</td>
<td>0.2744</td>
<td>0.3738</td>
<td>0.7921</td>
</tr>
<tr>
<td>MedicalCondition</td>
<td>0.0290</td>
<td>0.2842</td>
<td>0.2868</td>
<td>0.3697</td>
</tr>
<tr>
<td>TestOrMeasure</td>
<td>0.1425</td>
<td>0.0680</td>
<td>0.0920</td>
<td>0.1914</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.3080</td>
<td>0.1494</td>
<td>0.2012</td>
<td>0.4688</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.2670</td>
<td>0.2283</td>
<td>0.2462</td>
<td>0.4050</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.3037</td>
<td>0.3057</td>
<td>0.3047</td>
<td>0.4207</td>
</tr>
</tbody>
</table>

Figure 3.6: F-score measurements of exact and relaxed matching on different concept type.

It is obviously more desirable for a semi-automatic workflow to approximate the gold standard annotations (i.e., to produce exact matches rather than partial ones). Nevertheless, Argo’s automatically generated annotations proved to be helpful in a number of cases. For example, the automatic workflow was able to correctly annotate partially correct annotations such as *sputum* (for *sputum smear*), *pulmonary*
CHAPTER 3. CORPUS DEVELOPMENT

(for pulmonary TB) and COPD-staging (for COPD) which served as visual cues to the annotator. Based on her experience in annotating our corpus, she reported that having pre-supplied annotations, albeit incomplete or incorrect, is preferable over not having any annotations at all. We are, however, aware of the potential bias that having pre-supplied annotations may bring about, i.e., failure to annotate concepts completely missed by automatic annotation due to reliance on visual cues. To avoid this scenario, the annotator has been asked to read all of the sentences thoroughly and to keep in mind that the cues are not to be relied on. She has adhered to this guideline throughout her annotations.

Using the gold standard annotations to an information extraction task, we employed NERsuite to develop a new set of concept recognisers. Samples were represented using features which are by default extracted by NERsuite, including character, token, lemma and part-of-speech tag n-grams (within a distance of 2 from the token under consideration), chunk tags, as well as a comprehensive set of orthographic features (e.g., presence of uppercase or lowercase letters, digits, special characters). The resulting models were then evaluated in two ways. Firstly, for each concept type, models were trained and subsequently evaluated in a 10-fold cross-validation manner, whose results are presented in Table 3.12 alongside those obtained by the Argo components and the MetaMap. As we only have 30 annotated papers and in order to generate reasonable amounts of training and test documents, in generating the folds, the articles were split at the paragraph level, giving a total of 1,299 shorter documents. Even though some paragraphs from the same paper would occur in both training and test sets, those splitted paragraphs are closer to the real-life situations than using the whole papers. Secondly, to facilitate evaluation on unseen data, each of the automatically and manually annotated subset of 30 papers was subdivided into training (75% or 974 paragraphs) and held-out data (25% or 325 paragraphs).

Models trained on the former were then evaluated using annotations contained in the latter. Table 3.13 presents the evaluation results under this setting. The F-scores of both evaluation methods are also visualised in Figures 3.7 and 3.8.
Table 3.12: Results of 10-fold cross validation of concept recognisers, using exact matching
(Performance is compared with that of the components utilised in the text mining-assisted workflow).

<table>
<thead>
<tr>
<th>Concept type</th>
<th>Concept recognisers currently in Argo</th>
<th>MetaMap</th>
<th>Concept recognisers trained on our corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
</tr>
<tr>
<td>AnatomicalConcept</td>
<td>0.2361</td>
<td>0.6617</td>
<td>0.3428</td>
</tr>
<tr>
<td>Drug</td>
<td>0.7318</td>
<td>0.2161</td>
<td>0.3283</td>
</tr>
<tr>
<td>MedicalCondition</td>
<td>0.3986</td>
<td>0.2436</td>
<td>0.3010</td>
</tr>
<tr>
<td>TestOrMeasure</td>
<td>0.0766</td>
<td>0.0182</td>
<td>0.0289</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.4330</td>
<td>0.1021</td>
<td>0.1635</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.3305</td>
<td>0.1776</td>
<td>0.2310</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.3752</td>
<td>0.2483</td>
<td>0.2988</td>
</tr>
</tbody>
</table>
Table 3.13: Results of evaluation using a fixed split over 1,299 paragraphs
(Training set: 75% or 974 paragraphs; held-out set: 25% or 325 paragraphs), using exact matching.

<table>
<thead>
<tr>
<th>Concept type</th>
<th>Concept recognisers currently in Argo</th>
<th>MetaMap</th>
<th>Concept recognisers trained on our corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
</tr>
<tr>
<td>AnatomicalConcept</td>
<td>0.2602</td>
<td>0.6145</td>
<td>0.3656</td>
</tr>
<tr>
<td>Drug</td>
<td>0.6885</td>
<td>0.1900</td>
<td>0.2979</td>
</tr>
<tr>
<td>MedicalCondition</td>
<td>0.4494</td>
<td>0.2492</td>
<td>0.3206</td>
</tr>
<tr>
<td>TestOrMeasure</td>
<td>0.0250</td>
<td>0.0041</td>
<td>0.0070</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.4111</td>
<td>0.0847</td>
<td>0.1404</td>
</tr>
<tr>
<td>Micro-average</td>
<td>0.3735</td>
<td>0.1614</td>
<td>0.2254</td>
</tr>
<tr>
<td>Macro-average</td>
<td>0.3669</td>
<td>0.2285</td>
<td>0.2816</td>
</tr>
</tbody>
</table>
Figure 3.7: Comparative evaluation results of 10-fold cross validation using concept recognisers trained on our corpus.

Figure 3.8: Comparative evaluation results of unseen data annotation using concept recognisers trained on our corpus.
We show that by using our gold standard annotations as training data, we were able to develop concept recognisers whose performance is drastically better than those we employed in our semi-automatic workflow. This improvement ranged from 24.84 (for AnatomicalConcept) to 40.43 (for TestOrMeasure) percentage points according to 10-fold cross validation, and from 19.49 (for AnatomicalConcept) to 40.45 (for TestOrMeasure) according to the fixed split evaluation. Compared to the results generated by MetaMap, the Micro-/Macro-average F-scores are improved from 0.28 to 0.50. This implies that our corpus can stimulate the development of more suitable automatic COPD phenotype extractors.

Apart from the probability of wrongly recognised or unrecognised entities, the false annotations can be explained by three cases. Firstly, some expressions such as certain treatments do not require decomposition into elementary concepts, e.g., dietcontrol and pressure support ventilation. Secondly, there are instances of phenotypic expressions which were only partially recognised. For instance, in the excerpt, “Abdomen was soft and nontender, nondistended”, the problems which were automatically annotated span the qualities (i.e., nontender, nondistended) but not the anatomical entity of interest (i.e., Abdomen). Lastly, the entities contained in some phenotypic expressions do not correspond to the concept types that our current tools can automatically recognise. Expressions such as a persistent air leak, increasing wheezes, decreased breath sounds, for example, are highly domain-specific and contain entities (e.g., air, wheezes, breath) which can only be identified by tailor-made recognisers.

3.5 Summary

In this chapter, we elucidated our proposed text mining-assisted methodology for the gold-standard annotation of COPD phenotypes. We demonstrated with the proposed scheme that the annotation task can be kept simple for curators whilst producing expressive and computable annotations. By constructing a semi-automatic annotation workflow in Argo, we seamlessly integrate and take advantage of several automatic NLP tools for the task. Furthermore, we are providing the domain experts with a highly intuitive interface for creating and manipulating annotations. New concept recognisers trained on these gold standard annotations demonstrate
dramatically better performance (i.e., with a 20- to 30-percentage point margin in terms of F-scores) over the off-the-shelf components used in the Argo workflow.
Chapter 4

Phenotype normalisation

Gathering and analysing detailed information about phenotypes is important in various different scenarios, e.g., to provide medical professionals with sufficient evidence to provide appropriate therapy to an individual patient (Han et al. 2010) or to help patients and their families to better understand their diseases and treatments (Suominen et al. 2013). For such purposes, it can be important to consider details from multiple sources, since they may provide complementary information. For example, whilst clinical records typically contain various types of information about individual patients, biomedical research articles often provide summaries of the latest research findings, results and advances in knowledge (Patrick et al. 2007; 2008).

However, a barrier to the efficient location of information about concepts of interest is the many potential ways in which each concept could be mentioned in text. Such variations may arise due to factors such as the different writing styles of individual authors, or to avoid repetition of the same phrases, or according to the conventions typically followed in the text type in question, etc. In particular, the
characteristics of concept mentions in biomedical academic text and narrative clinical records are often very different. For example, the formal and technical nature of the language used in biomedical articles results in a fairly high proportion of concept mentions which are derived from the Greek language (e.g., hypercholesterolemia, dyspnea, hypertension, or leukocytosis). In contrast, the more informal style of information in clinical records results in the more dominant use of brief narrative phrases to describe concepts (e.g., cholesterol elevation, could not breathe, or blood pressure is high), along with a high occurrence of acronyms and abbreviations (e.g., high WBC, SOB, or HTN). Accordingly, a single concept may be expressed in text using phrases with different internal structures, and which bear no resemblance to each other (e.g., hypertension vs. blood pressure is high). Thus, particularly when searching across heterogeneous text types, it is very hard for a user to formulate search queries that account for all possible ways in which each concept of interest could be mentioned in text. This situation can inevitably lead to vital information being overlooked.

Automatic normalisation methods aim to determine the unique concept that is referred to by individual phrases in text. Such methods can ultimately facilitate the development of search systems that locate all mentions of a concept across large document collections, without the user having to try to enumerate each of the potentially diverse ways in which the concept may be expressed in the documents. Usually, normalisation methods work by trying to map each concept mention occurring in text to an entry in a domain-specific terminological resource. Typically, each entry in such a resource corresponds to a unique concept, and includes a limited number of manually-curated concept synonyms, which reflect some of the ways in which the concept could be mentioned in text. Normalisation methods generally perform mapping by trying to pair each concept mention in text to the most appropriate concept synonym in the terminological resource. This is often a challenging task, largely due to the creative nature of language usage. Since the list of concept synonyms provided in terminological resources is not intended to be exhaustive, there will often be no exact match between a concept mention in text and the most appropriate synonym in the terminological resource. Thus, the goal of normalisation methods is to find the optimal way to bridge the gap between concept mentions occurring in text and concept synonyms listed in terminological resources. The complexity of this challenge is intensified when considering text from
heterogeneous domains, due to the differing styles/conventions of describing concepts.

In this chapter, we describe the development of a novel, hybrid normalisation method, which is specifically intended for cross-domain application, and whose design has been guided by a detailed analysis of a large number of samples of clinical narrative text and biomedical academic text, together with concept mentions extracted from them, using text mining methods as we discussed in Chapter 2 and 3 (Fu and Ananiadou, 2014; Fu et al., 2014, 2015). Based on this analysis, we identified several important categories of concept variations that can occur both within and across domains, and we have developed a number of individual techniques that are tailored to handle these different types of variation, e.g., kidney disease vs. nephrosis, lung volume reduction vs. reduced lung volume, WBC vs. white blood cell and high blood sugars vs. high blood sugar. We have taken particular care to account for common cross-domain differences in concept mentions, e.g., to allow links to be made between the Greek-derived terms that are common in biomedical articles and the more descriptive phrases that are characteristic of narrative clinical text. We also address the ambiguous nature of acronyms/abbreviations, which can be particularly prevalent in clinical text.

We have investigated how to combine the individual techniques in an optimal manner, to create a hybrid method that achieves robust normalisation performance when applied to a wide range of concept mentions occurring in both biomedical abstracts and narrative clinical text. Through evaluation on gold standard annotated corpora for both domains, we show that our hybrid methods offer considerable performance improvements over a dictionary-based baseline, in which textual concept mentions are matched exactly to concept synonyms in a terminological resource. We additionally demonstrate the flexible nature of our method by showing that normalisation can be carried out successfully to entries in different terminological resources.
4.1 Literature review

4.1.1 Concept normalisation

A range of normalisation methods have proposed heuristics to non-exact matches between concept mentions in text and concept synonyms in terminological resources. These often apply fairly systematic transformations of concept mentions, such as removing inflections (e.g. plurals), ignoring certain non-matching words, generating derivations of words (e.g., *elevate* -> *elevation*), or generating permutations of words in concept synonyms (e.g., *increase in blood pressure* -> *blood pressure increase*) (Hersh and Greenes 1990; Jonquet, Shah, and Musen 2009; Savova et al. 2010). String similarity methods, which assign a numerical score representing the degree of similarity between an entity mention and a concept synonym, have also been applied as a more flexible means of determining approximate matches (*Doğan* and Lu 2012; *Kate* 2016). MetaMap is a very mature tool to generate morphological, lexical and syntactic variants, which has employed a variety of heuristics to try to account for the potential multitude of ways in which concepts can be mentioned in text. *Jacquemin* (1999; 1994) exploited knowledge of such variants, and particularly of syntactic transformations, in his FASTR system, which was developed for the recognition of term variants in textual corpora and has been successfully applied to such languages as French and English (*Jacquemin* 1996; *Jacquemin* et al. 1997).

A particular drawback of all of the above approaches is their reliance on finding surface-level similarities between concept mentions in text and one of the typically incomplete set of synonyms listed in the underlying terminological resource. Often, however, there can be various semantic-level variations amongst mentions of a concept (i.e., whose surface-level forms bear no resemblance to each other), which can be particularly prevalent when considering documents from heterogeneous domains. If any such synonyms are not listed in the terminological resource, then normalisation of these mentions will fail. Further difficulties in normalisation can arise from the considerable variation in the internal, syntactic structure of concept mentions, which are particularly frequent in narrative clinical text, and which can vary from simple noun phrases to complete clauses.

Over the past few years, a small number of gold standard annotated corpora have been released (*Alnazzawi* et al., 2016; *Wang* et al. 2016; *Doğan* et al., 2014;
Suominen et al. 2013; Leaman and Miller 2009), in which domain experts have manually mapped concept mentions occurring in text belonging to different domains (mainly biomedical abstracts and narrative clinical reports) to appropriate concept entries in domain specific terminological resources (such as MeSH, SNOMED-CT, and the UMLS Metathesaurus). The annotated corpora provide important evidence needed to develop accurate normalisation methods (i.e., examples of the range of ways in which concept mentions can vary in “real” text), as well as providing the means to evaluate the performance of different normalisation methods, by comparing their output to the expert-added annotations. Especially since some of these corpora were released in the context of shared tasks, their emergence has resulted in a recent surge in interest to develop novel approaches to normalisation for application to biomedical or clinical text. However, since the majority of annotated corpora and shared tasks have only been concerned with text from a single domain (either biomedical or clinical), most methods are specialised for application in one of these domains.

In general, the highest performing solutions have been those based on supervised machine-learning methods (Leaman et al. 2013). However, they are usually highly sensitive to the domain/text type on which they were trained. The restricted number of available suitably annotated corpora, combined with the time and expense required to create new corpora, can limit the feasibility of this approach in developing normalisation solutions that can perform robustly across different text types/domains. Other promising approaches have combined/ranked the results obtained through matching mentions to a number of different terminological resources (Collier et al. 2015), or used pattern-matching or regular expressions, which can account for frequently occurring variations not listed in the terminological resource (e.g., Greek or Roman suffixes for genes) (Fan et al. 2013; Ramanan et al. 2013; Wang and Akella 2013). The most similar work to ours is the PhenoNorm method reported in Alnazzawi et al. (2016), where the use of a string similarity method, combined with information from a general language lexical resource to facilitate generation of synonyms, was shown to constitute a flexible means to account for many surface-level and semantic-level differences amongst concept mentions in both clinical and biomedical text. However, the fact that string similarity methods calculate the degree of surface-level similarity between phrases, rather than considering the compositional similarity of the phrases being compared,
resulted in some mapping errors, such as *increased oxygen requirement* being mapped to *increased insulin requirement*.

### 4.1.2 Domain adaptation

Due to obvious variations in concept expressions from biomedical and clinical domains, several domain adaptation (DA) methods have been raised to bridge the gap between biomedical and clinical domains. DA is the adaption of a model trained on one domain for use in another domain whose distribution is related, but not identical to the original one (Daumé et al. 2010).

#### 4.1.2.1 Structural correspondence learning

Blitzer et al. (2007; 2006) introduced a semi-supervised method for DA called structural correspondence learning (SCL) which applied pivot features that occur frequently and behave in the same way in both domains to identify correspondences among features from different domains. Non-pivot features from those domains which are correlated with the same pivot features are supposed to correspond. They used SCL to transfer a POS tagger from the Wall Street Journal to MEDLINE. SCL consistently outperformed both supervised and semi-supervised learning with no labelled target domain training data. They also provided a refined version of SCL algorithm and used it for sentiment classification on a corpus of reviews for four different types of products from Amazon (Blitzer et al. 2007). Apart from feature frequency, they added mutual information to select pivot features. Then in order to achieve low target domain error, the $A-$distance (Ben-David et al. 2007) has been utilised to select the source domains which are more similar to the target domain. As a result, the new SCL algorithm reduced the relative error due to adaptation between domains by an average of 30% over the original SCL method and 46% over a supervised baseline.

SCL is the first method that uses unlabeled data from both domains and makes better use of the small amount of unlabeled data than those previous approaches. It focuses on finding a common representation for features from different domains instead of instances, because the same instance can contain some features that are common across domains while some others that are not. However, SCL assumes that the data from different domains are represented by the same type of features with the same dimension, and thus they are unable to deal with the problem where the
dimensions of data from the source and target domains are different, which is known as heterogeneous domain adaptation (Bel et al. 2003; Li et al. 2013).

### 4.1.2.2 Feature augmentation

Daumé et al. (2006; 2007) proposed a simple fully supervised approach to solve the DA problem in sequence labeling tasks, including named entity recognition, shallow parsing and part-of-speech tagging. Features of the corpora were divided into three groups, i.e. general features, source-specific features and target-specific features, so that the feature space was augmented. Classifiers can then be learnt by optimising feature-augmented weight vectors simultaneously. Besides, this technique can be easily extended to K-domain adaptation problem by making K+1 copies of the original feature space. In the work of Daumé et al. (2010), unlabeled data in the target domain, which is generally available in practical NLP tasks, were leveraged by another feature map to improve the performance of the former method, if the source and target domains are reasonably close to each other.

This approach to DA is very simple to implement and can be applied as a pre-processing step to any supervised learner. Nevertheless, Daumé et al. considered the adaptation as merely augmenting the feature space and each feature has the same prior mean and variance, regardless of whether it is general or domain specific, which would limit the performance of this approach.

### 4.1.2.3 Hierarchical Bayesian domain adaptation

To resolve the aforementioned problem, another study was done by Finkel and Manning (2009), who considered their directed Hierarchical Bayesian Domain Adaptation (HBDA) model as a formal version of the feature augmentation method (Daumé 2007). Comparing their model with the one proposed by Daumé, three modifications were made. Firstly, Finkel and Manning focused on the field of multi-task learning where the goal is to improve model performance across all tasks rather than on the target data only. Secondly, they assigned different values for variances of each domain and the domain-independent parameters, i.e. hyper-parameters, which proved useful to significantly improve model performance, while each feature of Daumé’s model had the same prior mean and variance. Finally, they discussed the influence of inconsistent label in train and test tasks on the performance of a CRF sequence model for named entity recognition and demonstrated that the HBDA
CHAPTER 4. PHENOTYPE NORMALISATION

model outperformed previous work and showed improvements on both named entity recognition and dependency parsing.

4.1.2.4 Instance weighting and stacking model

More recently, Miwa et al. (2012) have applied two DA methods in bio-event extraction. If the event types and possible arguments of source corpora acted in the same way as those in the target corpus, instance weighting method (Jiang and Zhai 2007) was adopted to add those event instances from the source corpora to the training set. The objective function was also modified to avoid skewed results. While there were only a small set of event types and arguments sharing between the source and the target corpus, the researchers used a stacking model that makes each module of event extraction be trained separately on the source corpus and incorporate the output as additional features when training the same module on the target domain. The use of those two DA methods resulted in improvements over the original system by 1.8% F-Score and 1.3%, respectively.

4.2 Methodology

Based on the above review, we can conclude that even though several domain adaptation methods have been proposed, there is still clear need for normalisation methods that can be robustly applied across texts belonging to multiple domains without adaptation, and which can better account for the observations that a) certain concept mentions may have forms that bear no resemblance to concept synonyms that are listed in terminological resources, and b) some concept mentions, such as acronyms and abbreviations, may be context dependent.

Our aim has thus been to develop a normalisation method that can fulfil these requirements, and which can be successfully applied across multiple domains, without the need to be adapted or retrained.

Based on the observed features of concept mentions in the clinical and biomedical domains, we take inspiration from a number of methods that have proven to be successful for normalisation, including those based on surface-level transformations, and those which generate variations using information from lexical resources.

In this section, we describe our approaches to handling the concept mention variation introduced above. In all cases, our techniques aim to generate variants of
concept mentions that appear in text, which are then matched against concept synonyms in a terminological resource. Taking into account the outlined weaknesses of the PhenoNorm method (Alnazzawi et al. 2016), our own methods aim to generate compositional synonyms.

4.2.1 Compositional synonym generation

We have developed two different approaches to generating compositional synonyms of concept mentions (i.e., variants whose surface form varies considerably or is completely different from the original concept mention), based on the use of terminological/lexical resources. While it was shown in Alnazzawi et al. (2016) that the use of information in the general language WordNet resource (Fellbaum 1998) could allow the generation of variants with appropriate synonyms of modifier words, e.g., elevated pulmonary capillary wedge pressure vs. increased pulmonary capillary wedge pressure, we make use of information in different specialised medical/biomedical resources to generate alternative types of variants.

4.2.1.1 Synonym searching

The UMLS Metathesaurus, a large and widely-used repository of biomedical terminology, usually includes a number of synonyms for each listed concept. In our approach, we firstly try to find exact matches between concept mentions in text and UMLS concept synonyms. We used the UMLS Application Programming Interface (API) to programmatically access the terminology. In the case of a match, all other synonyms of the concept are retrieved and added as potential variants of the mention. So, for example, through lookup in the UMLS Metathesaurus, Hypercholesterolemia and Hypotension can be found as variants of the textual concept mentions elevated cholesterol and blood pressure is low, respectively. Figure 4.1 is a screen shot of the search results through UMLS Terminology Services (UTS) Metathesaurus Browser.

However, a large number of concept mentions consist of multiple words. Given that the individual words occurring within such mentions may each have multiple synonyms, it is highly unlikely that all possible textual variants of the complete concept will be listed in UMLS. For example, the concept mention high WBC cannot be matched exactly with any concept synonyms in UMLS. Thus, if the expression is only considered as a whole, then no potential variants can be retrieved.
In such situations, each word in the expression is separately looked up in the Metathesaurus, and any listed synonyms are combined to generate potential variants. In this case, high is kept in the original form while White Blood Cell Count procedure is found as a synonym for WBC. These individual variants are then combined to generate a potential variant for the whole term, i.e., high white blood cell count procedure (as shown in Figure 4.2).

Figure 4.1: Screenshots of the example of the search results of elevated cholesterol and blood pressure is low through UTS Metathesaurus Browser
(a) Hypercholesterolemia is returned for elevated cholesterol;
(b) Hypotension is returned for blood pressure is low.
Figure 4.2: Screenshots of the example of the search results of *high WBC* through UTS Metathesaurus Browser

(a) *High WBC* cannot be found in the terminology;
(b) *High* is returned for *high*;
(c) *White Blood Cell Count procedure* is returned for *WBC*.
4.2.1.2 Transformation between English and Greek phenotypes

Greek (or the Hippocratic writings) is the oldest written sources of western medicine, covering all aspects of medicine and containing numerous medical terms. Medical Greek has been used since the 5th and 4th centuries BC, which was considered to be the beginning of the Greek era of the language of medicine (Wulff 2004). Latin was then introduced in 1478 and widely used to present the medical terms. Nevertheless, most of the new concepts developed by medical scientists were composed of Greek rather than Latin roots, as Latin does not to the same extent permit the formation of composite words (Wulff 2004). For example, the Greek terms nephrectomy, ophthalmoscopy and erythrocyte were introduced instead of those more cumbersome terms in medical Latin, i.e., excisio renis, inspectio oculorum and cellula rubra.

This huge term stock in Greek also presents the characteristics of linguistic interest such as the special meaning attached to certain prefixes (e.g. “rhachi-” (spine)), suffixes (“-itis” (inflammation)), or compound elements (e.g., “mast-” (breast) and “-omata” (tumour, mass, fluid collection)) of a Greek origin. Another fact is that some prefixes and suffixes are more productive than others. For example, Greek prefix “hyper-” is more productive than Latin “super-”, despite the fact that they originally have exactly the same meaning. As a result, hypertension that begins with the Greek prefix is used rather than supertension.

Nowadays, as English has become the language of choice of medical scientists for international communication, more and more new medical terms are composed of words from ordinary English, such as bypass operation, clearance, screening, and scanning rather than being derived from classical Greek or Latin roots. Markó et al. (2005) decomposed complex words into sub-words with equivalences to enhance cross-language retrieval performance, and they are also related to the more linguistically grounded methods of morphological analysis of medical neoclassical compounds (Deléger et al. 2007; Namer and Baud 2005; Ananiadou 1988).

Our second approach to the generation of variants addresses our observation that concept mentions in formal text are often derived from the Greek language, whilst mentions in clinical text are more likely to be corresponding to English phrases having roughly equivalent meanings to the Greek-based terms (some examples are displayed in Table 4.1). However, it is not always the case that the terminological resource used for normalisation will include both Greek terms and
their more descriptive English equivalents. For example, SNOMED-CT does not include any descriptive synonyms for the Greek-derived term *asthenia*. To address this, we have developed methods to generate English equivalents of Greek terms, and vice versa, in order to increase the likelihood of resolving concept mentions in text to appropriate concepts in the terminological resource.

Table 4.1: Examples of phenotypic concepts expressed in Greek and English.

<table>
<thead>
<tr>
<th>Greek concept in biomedical literature</th>
<th>English concept with the same meaning in clinical records</th>
</tr>
</thead>
<tbody>
<tr>
<td>aphasia</td>
<td>language dysfunction</td>
</tr>
<tr>
<td>apnea</td>
<td>not breathing</td>
</tr>
<tr>
<td>dysphagia</td>
<td>failed swallow</td>
</tr>
<tr>
<td>dyspnea</td>
<td>shortness of breath</td>
</tr>
<tr>
<td>hyperglycemia</td>
<td>elevated blood sugar</td>
</tr>
<tr>
<td>hypertension</td>
<td>high blood pressure</td>
</tr>
<tr>
<td>thyromegaly</td>
<td>thyroid enlarged</td>
</tr>
</tbody>
</table>

Our methods make use of a list of roots, suffixes, and prefixes commonly employed in medical terminology, and which are derived from ancient Greek, which contains 809 Greek elements (i.e., roots or affixes) and their equivalents in English. Some examples are shown in Table 4.2.

In this list, each Greek root or affix is accompanied by one or more English equivalents. For concept mentions corresponding to multiple words of English origin, Greek elements corresponding to each English word are retrieved from the list and are combined or composed into a potential single-word Greek term. The procedure is shown in Figure 4.3. For *difficult breathing*, “dys-” and “-pne-, -pnoea” are first retrieved from the medical element list for the two English words “difficult” and “breathing”, separately, and then are combined to create the new phenotypes in Greek, i.e., *dyspnea* and *dyspnoea*.

---

Table 4.2: Examples of Greek elements and their expression in English.

<table>
<thead>
<tr>
<th>Greek element</th>
<th>Equivalent English expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>abdomin-/ abdomino-</td>
<td>abdomen, abdominal</td>
</tr>
<tr>
<td>cyt-/ cyto-/ -cyte</td>
<td>cell</td>
</tr>
<tr>
<td>-dynia</td>
<td>pain</td>
</tr>
<tr>
<td>-esophageal/ -esophago</td>
<td>gullet</td>
</tr>
<tr>
<td>gastr-/ gastro-</td>
<td>stomach</td>
</tr>
<tr>
<td>hepat-/ hepatic-</td>
<td>liver</td>
</tr>
<tr>
<td>-ismus</td>
<td>spasm, contraction</td>
</tr>
<tr>
<td>kin-/ kine-/ kino-/ kinesi-/ kinesio-</td>
<td>movement</td>
</tr>
<tr>
<td>leuc-/ leuco-/ leuk-/ leuko-</td>
<td>white</td>
</tr>
<tr>
<td>nerv-/ neur-/ neuri-/ neuro-</td>
<td>nerves, nervous system</td>
</tr>
<tr>
<td>olig-/ oligo-</td>
<td>little, few</td>
</tr>
<tr>
<td>xanth-/ xantho-</td>
<td>yellow, abnormally yellow</td>
</tr>
</tbody>
</table>

The list of medical elements

<table>
<thead>
<tr>
<th>English expression</th>
<th>Greek element</th>
</tr>
</thead>
<tbody>
<tr>
<td>difficult breathing</td>
<td>dys-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Figure 4.3: Example of the formation of Greek phenotypes.

Conversely, concept mentions that correspond to Greek terms usually consist of a single word, which can often be decomposed into several parts, each of which may be “translated” using one or more English words, according to the information in the
list. In this case, according to the number of possible English translations, a number of potential English multi-word variants many be generated. As depicted in Figure 4.4, the single-word concept (e.g., *dyspnea*) will be decomposed into several parts (e.g., “dys-” and “-pnea”) whose corresponding meanings in English (e.g., “difficult, abnormal, failed, etc.” and “breath, breathing”) will be retrieved from the list to form new phrases (e.g., *difficult breath, difficult breathing, abnormal breath, abnormal breathing*, etc.).

The list of medical elements

<table>
<thead>
<tr>
<th>Greek element</th>
<th>English expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>dys-</td>
<td>difficult, abnormal, defective, bad, failed, difficulty</td>
</tr>
<tr>
<td>-pnea</td>
<td>breath, breathing</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Some examples of composition and decomposition between English and Greek concepts are shown in Table 4.3. For instance, a potential English equivalent for the Greek prefix *dys-* is *failed* while the Greek suffix *-phagia* is connected with eating and swallowing. Therefore, the Greek variant of the concept mention *failed swallow* is *dysphagia*. Conversely, the Greek term *asthenia* can be decomposed into *a-*(without) and *-sthenia* (strength or power). Accordingly, we generate the English equivalents of this term *without strength* or *without power*. Furthermore, since *asthenia* as a whole could also be translated as *weakness* or *fatigue*, we also generate these as variants. The above process of equivalence generation between Greek and English is fully automatic without any human validation.
Table 4.3: Examples of transformation between English and Greek concepts
(a) English to Greek; (b) Greek to English.

<table>
<thead>
<tr>
<th>Original English concept</th>
<th>Corresponding elements</th>
<th>Generated Greek concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>failed swallow</td>
<td>failed: dys-;</td>
<td>dysphagia</td>
</tr>
<tr>
<td></td>
<td>swallow: -phagia.</td>
<td></td>
</tr>
<tr>
<td>high blood sugar</td>
<td>high: hyper-;</td>
<td>hyperglycemia</td>
</tr>
<tr>
<td></td>
<td>blood: -emia;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sugar: glyc-.</td>
<td></td>
</tr>
<tr>
<td>without strength</td>
<td>without: a-;</td>
<td>asthenia</td>
</tr>
<tr>
<td></td>
<td>strength: -stenia.</td>
<td></td>
</tr>
<tr>
<td>thyroid enlargement</td>
<td>thyroid: thyro-;</td>
<td>thyromegaly</td>
</tr>
<tr>
<td></td>
<td>enlargement: -megaly.</td>
<td></td>
</tr>
<tr>
<td>swollen tonsils</td>
<td>swollen: -itis;</td>
<td>tonsillitis</td>
</tr>
<tr>
<td></td>
<td>tonsils: tonsil-.</td>
<td></td>
</tr>
<tr>
<td>increased white blood cells</td>
<td>white, white blood cells: leuko-;</td>
<td>leukocytosis</td>
</tr>
<tr>
<td></td>
<td>cell: cyt-;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>increased: -osis.</td>
<td></td>
</tr>
</tbody>
</table>

(a)
<table>
<thead>
<tr>
<th>Original Greek concept</th>
<th>Elements and their meanings</th>
<th>Generated English concept(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dysphagia</td>
<td>dys-: failed, difficult;</td>
<td>failed swallow(ing), failed eat(ing), difficult swallow(ing)…</td>
</tr>
<tr>
<td></td>
<td>-phagia: eat(ing), swallow(ing).</td>
<td></td>
</tr>
<tr>
<td>hyperglycemia</td>
<td>hyper-: (abnormally) high, elevated;</td>
<td>high sugar blood, high blood sugar, elevated blood sugar…</td>
</tr>
<tr>
<td></td>
<td>glyc-: sweet, sugar;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-emia: blood (condition).</td>
<td></td>
</tr>
<tr>
<td>asthenia</td>
<td>a-: without, loss of;</td>
<td>without strength, loss of strength, weakness, fatigue…</td>
</tr>
<tr>
<td></td>
<td>-sthenia: strength, power;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-asthenia: weakness, fatigue.</td>
<td></td>
</tr>
<tr>
<td>hypoadrenalism</td>
<td>hypo-: insufficiency, below normal;</td>
<td>adrenal insufficiency (disease, condition), adrenal below normal (disease, condition)</td>
</tr>
<tr>
<td></td>
<td>-ism: disease, condition.</td>
<td></td>
</tr>
<tr>
<td>thyromegaly</td>
<td>thyro-: thyroid;</td>
<td>thyroid enlargement</td>
</tr>
<tr>
<td></td>
<td>megaly: enlargement.</td>
<td></td>
</tr>
<tr>
<td>tonsillitis</td>
<td>tonsill-: tonsil;</td>
<td>tonsil inflammation, tonsil swelling</td>
</tr>
<tr>
<td></td>
<td>-itis: inflammation, swelling.</td>
<td></td>
</tr>
<tr>
<td>leukocytosis</td>
<td>leuko-: white, white blood cells;</td>
<td>increased white blood cells (disease, condition)</td>
</tr>
<tr>
<td></td>
<td>cyt-: cell;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-osis: increased, disease, condition.</td>
<td></td>
</tr>
</tbody>
</table>
4.2.2 Normalisation of syntactic structures

4.2.2.1 Predicate-argument structures

As we have previously observed, differences in the internal syntactic structure of concept mentions constitute an important type of variation amongst mentions of the same concept. Changes in internal structure often also involve the use of different grammatically derived words, resulting in potential variations such as *cholesterol elevation, elevation of cholesterol, elevated cholesterol, and cholesterol elevated*; or *dyspnoea improvement, improvement in dyspnoea, improved dyspnoea, and dyspnoea improved*.

We take a systematic, syntactic approach to generating such variants. Firstly, we apply the probabilistic head-driven phrase structure grammar (HPSG) parser Enju⁶ (Miyao and Tsujii, 2008) to obtain the predicate-argument structures (PASes) of concepts, and thus to capture the relationships between head and dependent words. Five different types of relationships derived from Enju are shown in Table 4.4.

A set of predicate-argument relations was generated by Enju, from which one can easily acquire relations among words without the burden of analyzing their deep-syntactic structure. Parsing examples are shown in Table 4.5 and Figure 4.5. Each line in the output represents a predicate-argument relation between two words. For instance, the second line in the first example (i.e., *cholesterol elevation*) indicates that there is an “ARG1 (logical subject)” relation between the predicate *elevation* and the argument *cholesterol*, while for the second example (i.e., *elevation of cholesterol*), *elevation* and *cholesterol* are related with *of* as “ARG1” and “ARG2” in the “prep_arg12”. Similarly, *cholesterol* is treated as “ARG1” in both the third (e.g., *elevated cholesterol*) and fourth examples (e.g., *elevating cholesterol*).

---

⁶ http://www.nactem.ac.uk/enju/
Table 4.4: Examples of different syntactic constructions of concepts identified by Enju.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Concept</th>
<th>Predicate</th>
<th>Argument 1</th>
<th>Argument 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>noun_arg1</td>
<td>cholesterol elevation</td>
<td>cholesterol</td>
<td>elevation</td>
<td>-</td>
</tr>
<tr>
<td>C₂</td>
<td>prep_arg12</td>
<td>elevation of cholesterol</td>
<td>of</td>
<td>elevation</td>
<td>cholesterol</td>
</tr>
<tr>
<td>C₃</td>
<td>adv_arg1</td>
<td>elevated cholesterol</td>
<td>elevated</td>
<td>cholesterol</td>
<td>-</td>
</tr>
<tr>
<td>C₄</td>
<td>verb_arg1</td>
<td>elevating cholesterol</td>
<td>elevating</td>
<td>cholesterol</td>
<td>-</td>
</tr>
<tr>
<td>C₅</td>
<td>special cases</td>
<td>blood pressure is low could not breathe</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4.5: Output of the Enju parser for cholesterol elevation, elevation of cholesterol, elevated cholesterol, elevating cholesterol.

<table>
<thead>
<tr>
<th>cholesterol elevation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
</tr>
<tr>
<td>cholesterol</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>elevation of cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
</tr>
<tr>
<td>of</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>elevated cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
</tr>
<tr>
<td>elevated</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>elevating cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
</tr>
<tr>
<td>elevating</td>
</tr>
</tbody>
</table>
Figure 4.5: PASes produced by Enju
(a) $C_1$ (cholesterol elevation); (b) $C_2$ (elevation of cholesterol);
(c) $C_3$ (elevated cholesterol); (d) $C_4$ (elevating cholesterol).
For the phenotypes comprising more than two words (for \( C_2 \), more than three words), such as \textit{reversible airflow limitation, management of stable COPD, elevated arterial carbon dioxide}, we will repeat the previous steps to find their nesting PASs sequentially, as shown in Figure 4.6. For example, \textit{airflow limitation} belongs to \( C_1 \) while \textit{reversible [airflow limitation]} is an instance in \( C_3 \). Similarly, \textit{stable COPD} is in group \( C_3 \), and in the meantime, \textit{management of [stable COPD]} is a preposition phrase belonging to \( C_2 \).

Figure 4.6: Nesting PASs of phenotypes

(a) Reversible airflow limitation; (b) Management of stable COPD; (c) Elevated arterial carbon dioxide.
4.2.2.2 Generation rules

If a concept is categorised in $C_1$, it will then be transformed to $C_2$, $C_3$ and $C_4$, separately. An example of the transformation of *cholesterol elevation* is shown in Figure 4.7 and more details about how to implement the conversion between two categories are shown in Figure 4.8.

![Figure 4.7: The syntactic transformation of cholesterol elevation.](image)

(a)

(b)
CHAPTER 4. PHENOTYPE NORMALISATION

Figure 4.8: The conversion process between pairs of PASes
(a) $C_1$ to $C_2$; (b) $C_1$ to $C_3$; (c) $C_1$ to $C_4$.

BioLexicon\(^7\) is a linguistic resource tailored for the biology domain and contains information on terminological verbs, derived forms of the terminological verbs, general English words frequently used in the biology domain and domain terms (Thompson et al. 2011). It was used in our research to generate POS variants (e.g., elevation (noun (N)), elevated (adjective (AJ)), and elevate (verb (V)), and different tenses of verbs (e.g., elevated (past tense (VVD)) and elevating (gerund/participle (VVG))).

Specifically, subcategorisation frames (SubCat Frame) in BioLexicon are used to find the most suitable prepositions for $C_2$. Figure 4.9 displays the possible complete grammatical frames that can occur with transcribe. The first column is the SubCat Frame where ARG1 and ARG2 denote Argument 1 and Argument 2 as in the Enju parser corresponding to the grammatical subject and object of the verb, possibly accompanied by a phrase beginning with a preposition (e.g., in, with, to or on), and the second and the third columns represent the probability score of each frame. The higher the score, the more likely the preposition will occur, given the entry term. For example, the preposition *in* (ARG1#ARG2#PP-in) will be chosen for *elevate* (e.g., elevation *in* cholesterol) or *decrease* (e.g., decrease *in* blood pressure). The preposition with the highest probability score for *treat* is *with* (ARG1#ARG2#PP-with) (e.g., treatment *with* antibodies) whilst it will be *expose to* rather than *expose on* or *expose in* as the preposition *to* (ARG1#ARG2#PP-to) has a

\(^7\) http://wiki.ilc.cnr.it/BootStrep/searchPanel.action
higher probability score (e.g., *exposure to tobacco smoke*). Besides, the preposition *of* will be used in most other cases, such as *symptoms of cough, blockage of airway, loss of appetite, management of stable COPD, or weakness of respiratory muscle*.

(a) Results of *elevate* (ARG1#ARG2#PP-in);
(b) Results of *decrease* (ARG1#ARG2#PP-in);
(c) Results of *treat* (ARG1#ARG2#PP-with);
(d) Results of *expose* (ARG1#ARG2#PP-to).

Figure 4.9: Screenshots of examples of subcategorisation frames (SubCat Frame) with the probability scores of prepositions in the BioLexicon.
If a concept occurs in any of the forms C₂, C₃ or C₄, then variants of the other two forms are automatically generated using the same rules. However, after discussing with the clinicians, some specific medical terms, like the names of drugs (e.g., phosphodiesterase-4 inhibitor, angiotensin receptor neprilysin inhibitor, or Beta2-adrenergic agonist), diseases and disorders (e.g., Addison’s disease, congestive heart failure, anoxic brain injury, renal failure, or hepatitis C cirrhosis), the original forms will be kept in the texts.

The special cases (C₅) will be manually transformed as there are numerous variations. Several examples are shown in Table 4.6. For instance, blood pressure is low will be converted to low blood pressure, and could not breathe, inability to eat, can’t work will be replaced by unable to breathe/eat/walk, separately.

<table>
<thead>
<tr>
<th>Original phenotype</th>
<th>Transformed phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood pressure is low</td>
<td>low blood pressure</td>
</tr>
<tr>
<td>could not breathe</td>
<td>unable to breath</td>
</tr>
<tr>
<td>inability to eat</td>
<td>unable to eat</td>
</tr>
<tr>
<td>can’t walk</td>
<td>unable to walk</td>
</tr>
</tbody>
</table>

### 4.2.3 Normalisation of acronyms and abbreviations

In both academic biomedical and narrative clinical text, acronyms and abbreviations are widely used. Some of those are unambiguous and has almost become a noun in its own right, such as AIDS (acquired immune deficiency syndrome), MRI (magnetic resonance imaging), and PCR (polymerase chain reaction), whereas many others are ambiguous and have several potential expansions (Carroll et al., 2012), presenting the difficulty that they have different meanings based on their contexts. For instance, ASA has 40 definitions, such as argininosuccinic aciduria, ascorbic acid, and American Society of Anesthesiologists. VSD may mean any of ventricular septal defect, virulent systemic disease, and vascular surface density, depending on the context. Accordingly, concept mentions corresponding to acronyms and abbreviations cannot be effectively normalised without some kind of disambiguation.
If directly mapping those acronyms and abbreviations in a medical knowledge base, no specific results will be returned.

Therefore, in order to correctly recognise them, a disambiguation process needs to be applied. For this purpose, we determined the most appropriate expansions for each acronym and abbreviation using AcroMine Disambiguator (see Section 2.3.3) which has previously been shown to achieve an F-score of 0.986 when applied to MEDLINE records.

4.2.4 Normalisation of plural and singular

A final important method that has been proven to be important in other normalisation efforts is the transformation of concept mentions appearing in the plural form into their singular equivalents. This is important, since terminological resources typically list synonyms of concepts only in the singular form and hence, concept mentions in plural form will usually not match with any concept synonyms. Thus, all concept mentions occurring in the plural form are replaced with their base (i.e., singular) forms by making use of the information output by the Enju parser. As shown in Table 4.7, for instance, the concept mentions thrombi, fasciculations, and high blood sugars are converted to thrombus, fasciculation, and high blood sugar, respectively.
Table 4.7: Base forms of *thrombi*, *fasciculations*, and *high blood sugars* parsed by Enju.
4.2.5 Individual and hybrid methods

It has previously been demonstrated that hybrid systems, which combine the results of a number of different methods, can achieve superior performance to the individual methods from which they are constructed (Kang et al., 2010; Uzuner et al., 2011). Accordingly, we created and evaluated a number of different hybrid methods, which combine two or more of the individual methods introduced above. In each case, the best performing individual method is firstly applied to generate an initial set of variants for each concept mention. If any of these variants can be matched to a concept synonym in the appropriate terminological resource, then the method terminates. If no, variants are generated by the next best performing method are generated, and a match is sought amongst these. The process continues until a match is found between a generated variant of the concept mention and a concept synonym in the terminological resource, or until all methods combined in the hybrid have been applied.

4.3 Results and discussion

4.3.1 Variants generation

To ensure the cross-domain applicability of our methods, we have evaluated their performance on two corpora, i.e., the ShARe/CLEF corpus and the BioCreative corpus, that include gold-standard normalisation annotations, and whose texts belong to different domains, i.e., clinical and biomedical domains.

- **ShARe/CLEF corpus (Clinical domain)**
  We have used the 299 clinical records (i.e., discharge summaries, electrocardiogram, echocardiogram, and radiology reports) released in the context of the ShARe/CLEF eHealth Evaluation Lab 2013 (Suominen et al. 2013), which are annotated with normalised disorder mentions. The number of gold standard, normalised phenotype annotations in this corpus is 11,167 (3,821 unique), and these annotated mentions are normalised to concept unique identifiers (CUIs) in SNOMED-CT for the ShARe/CLEF corpus.
• **BioCreative corpus (Biomedical domain)**

A collection of 1,500 PubMed abstracts with the annotations of normalised disease mentions, which were used in the BioCreative V Chemical Disease Relation (CDR) Task (Wei et al. 2015), are also utilised in our research. In the BioCreative corpus, the number of gold standard, normalised phenotype annotations is 18,797 (3,737 unique), which are normalised to the MeSH IDs.

The annotated mentions in these two corpora are normalised to entries in two different terminological resources. This is also relevant for the evaluation of our method, since a flexible normalisation method should additionally facilitate accurate normalisation to concept entries in different resources.

Table 4.8 shows the total number of variants generated by applying each of our individual methods to all annotated concept mentions in each of the corpora. In some cases, a method may generate more than one variant for a given concept mention, while in others, there may be no variants generated for the concept mention. The transformation from Greek terms to English equivalents produced the greatest number of variants. This is expected, since each Greek element can typically be translated using several English words. All these normalisation methods were not applied to the terminology terms.

Table 4.8: Number of unique variants generated by different normalisation methods.

<table>
<thead>
<tr>
<th>Normalisation method</th>
<th>No. of unique variants generated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical</td>
</tr>
<tr>
<td>Synonym searching (SS)</td>
<td>545</td>
</tr>
<tr>
<td>English to Greek (E2G)</td>
<td>591</td>
</tr>
<tr>
<td>Greek to English (G2E)</td>
<td>2,060</td>
</tr>
<tr>
<td>Syntactic normalisation (SN)</td>
<td>1,411</td>
</tr>
<tr>
<td>Acronym and abbreviation disambiguation (AAD)</td>
<td>273</td>
</tr>
<tr>
<td>Plural to singular (P2S)</td>
<td>438</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5,318</strong></td>
</tr>
</tbody>
</table>
4.3.2 Evaluation

To calculate the performance of our methods in terms of F-score, we determined true positive (TP), false positive (FP) and false negative (FN) as follows:

1) Each case of a correct normalisation is counted as a TP, which means a normalised phenotype can be found in the standard terminologies (i.e., SNOMED-CT and MeSH) and are assigned a correct ID (i.e., CUIs or MeSH Unique IDs).

2) Each normalised phenotype that can be found in the terminologies but is mapped to an incorrect concept (or are assigned a wrong ID according to the gold-standard annotations) is counted as an FP. For example, the Unique ID of *tetanus* found in MeSH by our methods is D013742, but the correct ID provided by the BioCreative corpus is D013746 which stands for *tetany*, while our methods assign the CUIs, i.e., C1444783 and C0008301, to *instability* and *choking*, respectively, which are both marked as “CUI-less” in the ShARe/CLEF corpus (More details can be found on page 142).

3) Each normalised phenotype that cannot be found in the terminologies but should have been assign an ID is counted as an FN. For example, *cytotoxic oedema within cerebral white matter* cannot be mapped to *Brain Edema* (D001929) in MeSH and no result can be returned for *left lower lobe collapse* (C0004144: *atelectasis*) in SNOMED-CT either.

4) Each normalised phenotype that cannot be identified in both terminologies and with an unidentified annotation, namely, “CUI-less” in SNOMED-CT and “-1” in MeSH, is counted as a TN.

4.3.2.1 Baseline methods

We firstly developed a straightforward baseline approach that found exact matches between textual concept mentions and concept synonyms listed in either SNOMED- CT or MeSH (according to the corpus used). This method achieved an F-score of 0.6850 for the ShARe/CLEF corpus and 0.8015 for the BioCreative corpus.
We also developed a baseline system using the MetaMap JAVA API to programmatically process the list of original concepts. The F-scores generated by MetaMap are 0.6916 and 0.8030 in the clinical and the biomedical domains, respectively.

### 4.3.2.2 Normalisation methods

In Table 4.9, those baseline results are compared with those obtained through the individual application of our six individual normalisation methods, and the same results can be visualised in Figure 4.10. As can be observed, all methods were able to improve the normalisation performance over the baseline for both corpora, with the exception of the P2S method when applied to the biomedical BioCreative corpus. The general success of all individual methods clearly illustrates the need to consider multiple types of variation to achieve better normalisation performance. The highest precision was achieved by AAD and P2S in the Clinical and Biomedical corpora, respectively, while the best recall in both domains are generated by SS. The disambiguation of acronyms and abbreviations (i.e., AAD) achieved the greatest performance increase over the baseline for both domains (with F-scores of 0.7096 and 0.8365 in the clinical and biomedical corpora, respectively), illustrating the frequent occurrence and potentially diverse interpretations of such concept mentions in both text types.

<table>
<thead>
<tr>
<th>Method</th>
<th>Clinical</th>
<th></th>
<th></th>
<th>Biomedical</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
</tr>
<tr>
<td>Straightforward baseline</td>
<td>0.7120</td>
<td>0.6601</td>
<td>0.6850</td>
<td>0.9727</td>
<td>0.6815</td>
<td>0.8015</td>
</tr>
<tr>
<td>MetaMap</td>
<td>0.7151</td>
<td>0.6766</td>
<td>0.6953</td>
<td>0.9069</td>
<td>0.7251</td>
<td>0.8030</td>
</tr>
<tr>
<td>SS</td>
<td>0.7076</td>
<td>0.6930</td>
<td>0.7001</td>
<td>0.9077</td>
<td>0.7514</td>
<td>0.8222</td>
</tr>
<tr>
<td>G2E</td>
<td>0.7208</td>
<td>0.6660</td>
<td>0.6923</td>
<td>0.9727</td>
<td>0.6816</td>
<td>0.8016</td>
</tr>
<tr>
<td>E2G</td>
<td>0.7112</td>
<td>0.6826</td>
<td>0.6966</td>
<td>0.9719</td>
<td>0.6822</td>
<td>0.8017</td>
</tr>
<tr>
<td>SN</td>
<td>0.7142</td>
<td>0.6794</td>
<td>0.6963</td>
<td>0.9723</td>
<td>0.6820</td>
<td>0.8017</td>
</tr>
<tr>
<td>AAD</td>
<td>0.7226</td>
<td>0.6911</td>
<td><strong>0.7096</strong></td>
<td>0.9703</td>
<td>0.7351</td>
<td><strong>0.8365</strong></td>
</tr>
<tr>
<td>P2S</td>
<td>0.7081</td>
<td>0.6664</td>
<td>0.6869</td>
<td>0.9748</td>
<td>0.6564</td>
<td>0.7845</td>
</tr>
</tbody>
</table>

Table 4.9: Normalisation performance of individual methods in both domains.
Figure 4.10: Normalisation performance using individual methods
(a) Clinical domain; (b) Biomedical domain.
To investigate the impact of combining multiple normalisation techniques into a hybrid method, we firstly evaluated the results of combining the best-performing AAD method with each of the other five methods. The results of these two-method combinations are shown in the upper parts of Table 4.10. In all cases, the combined results are superior to the results of the individual methods, which demonstrates the complementary nature of the variants generated by the different methods. In both corpora, the greatest improvement over AAD was achieved through the addition of synonym variations (the SS method). This is unsurprising, given that SS achieved the second best results when applied individually, and emphasises the importance of considering semantic-level variants as well as surface-level variations. The remaining four methods (i.e., E2G, G2E, SN and P2S) were then added in the order of their corpus-specific normalisation performance, from highest to lowest. For example, the concept mention *renal diseases* does not match with a concept synonym in either SNOMED-CT or MeSH. However, through the retrieval of the synonym *kidney diseases* from the UMLS Metathesaurus (using the SS method), mappings to the correct SNOMED-CT CUI (C0022658) and MeSH ID (D007674) can be achieved. Although the E2G method generates a further plausible variant (i.e., *nephrosis*), this is actually a more specific term, and this corresponds to a separate concept entry in both SNOMED-CT (C0027720) and MeSH (D007674). Thus, E2G should only be applied if SS fails to achieve a match, which fits in with its ranking as the third best normalisation method. In the lower half of Table 4.8, we show how the incremental addition of individual methods in the order of their performance results in a continual improvement of the overall normalisation performance, reaching 0.7208 F-score for clinical text and 0.8908 for biomedical text. These represent performance increases of 5.23% and 11.14% over the straightforward baselines for each text type, respectively. While comparing to the MetaMap method, an increase of 3.67% and 10.19% in F-scores are achieved. All these results are also visualised in Figures 4.11 and 4.12 (scores).
Table 4.10: Normalisation performance of hybrid methods in both domains
(a) Clinical domain; (b) Biomedical domain (in the ascending order of F-scores).

(a)

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAD + G2E</td>
<td>0.7237</td>
<td>0.7024</td>
<td>0.7118</td>
</tr>
<tr>
<td>AAD + SN</td>
<td>0.7232</td>
<td>0.7019</td>
<td>0.7124</td>
</tr>
<tr>
<td>AAD + P2S</td>
<td>0.7250</td>
<td>0.7126</td>
<td>0.7188</td>
</tr>
<tr>
<td>AAD + E2G</td>
<td>0.7248</td>
<td>0.7131</td>
<td>0.7189</td>
</tr>
<tr>
<td>AAD + SS</td>
<td>0.7262</td>
<td>0.7126</td>
<td>0.7194</td>
</tr>
<tr>
<td>AAD + SS + E2G</td>
<td>0.7238</td>
<td>0.7153</td>
<td>0.7195</td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S</td>
<td>0.7233</td>
<td>0.7164</td>
<td>0.7198</td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S + SN</td>
<td>0.7232</td>
<td>0.7174</td>
<td>0.7203</td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S + SN + G2E</td>
<td>0.7232</td>
<td>0.7184</td>
<td>0.7208</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAD + G2E</td>
<td>0.9698</td>
<td>0.7351</td>
<td>0.8363</td>
</tr>
<tr>
<td>AAD + E2G</td>
<td>0.9694</td>
<td>0.7359</td>
<td>0.8366</td>
</tr>
<tr>
<td>AAD + SN</td>
<td>0.9710</td>
<td>0.7433</td>
<td>0.8420</td>
</tr>
<tr>
<td>AAD + P2S</td>
<td>0.9704</td>
<td>0.7535</td>
<td>0.8483</td>
</tr>
<tr>
<td>AAD + SS</td>
<td>0.9535</td>
<td>0.8231</td>
<td>0.8835</td>
</tr>
<tr>
<td>AAD + SS + P2S</td>
<td>0.9538</td>
<td>0.8353</td>
<td>0.8906</td>
</tr>
<tr>
<td>AAD + SS + P2S + SN</td>
<td>0.9537</td>
<td>0.8353</td>
<td>0.8906</td>
</tr>
<tr>
<td>AAD + SS + P2S + SN + E2G</td>
<td>0.9537</td>
<td>0.8355</td>
<td>0.8907</td>
</tr>
<tr>
<td>AAD + SS + P2S + SN + E2G + G2E</td>
<td>0.9535</td>
<td>0.8358</td>
<td>0.8908</td>
</tr>
</tbody>
</table>
Figure 4.11: Normalisation performance of hybrid methods in clinical domain
(a) Displayed with y axis from 0 to 1;
(b) In order to show the differences more clearly, the same diagram is enlarged by using y axis from 0.65 to 0.75.
Figure 4.12: Normalisation performance of hybrid methods in biomedical domain

(a) Displayed with y axis from 0 to 1;
(b) In order to show the differences more clearly, the same diagram is enlarged by using y axis from 0.65 to 1.
Our methods are also evaluated in terms of accuracy, which is defined in Equation 4.1.

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4.1)
\]

Since TP + TN represent the number of normalised entity mentions that are mapped to the correct concept or to no concept (correct), and the summation of TP, TN, FP, and FN is the total number of entity mentions in the corpus (total), accuracy can also be formulated as follows (Equation 4.2):

\[
\text{Accuracy} = \frac{\text{correct}}{\text{total}} \quad (4.2)
\]

The accuracy of our normalisation methods applied in the clinical and biomedical domains is shown in Table 4.11, and the same results are also visualised in Figure 4.13. All the methods achieved a higher accuracy for the biomedical literature than for the clinical records, and the highest accuracy (i.e., 0.7166 for the clinical domain and 0.8132 for the biomedical domain), is also obtained by the hybrid method of AAD, SS, E2G, P2S, SN, and G2E, which is the same as being evaluated using F-score. An increase of 12.81% and 20.92% over the MetaMap baselines are achieved for the clinical and biomedical texts, respectively.
### Table 4.11: Normalisation performance of all our methods in terms of accuracy in both domains

<table>
<thead>
<tr>
<th>Normalisation method</th>
<th>Accuracy</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical</td>
<td>Biomedical</td>
<td></td>
</tr>
<tr>
<td>Straightforward baseline</td>
<td>0.5505</td>
<td>0.6702</td>
<td></td>
</tr>
<tr>
<td>MetaMap</td>
<td>0.6352</td>
<td>0.6725</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>0.5926</td>
<td>0.6993</td>
<td></td>
</tr>
<tr>
<td>G2E</td>
<td>0.6045</td>
<td>0.6703</td>
<td></td>
</tr>
<tr>
<td>E2G</td>
<td>0.6511</td>
<td>0.6704</td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>0.6439</td>
<td>0.6707</td>
<td></td>
</tr>
<tr>
<td>AAD</td>
<td>0.6585</td>
<td>0.7201</td>
<td></td>
</tr>
<tr>
<td>P2S</td>
<td>0.6357</td>
<td>0.6470</td>
<td></td>
</tr>
<tr>
<td>AAD + G2E</td>
<td>0.6937</td>
<td>0.7202</td>
<td></td>
</tr>
<tr>
<td>AAD + SN</td>
<td>0.6923</td>
<td>0.7203</td>
<td></td>
</tr>
<tr>
<td>AAD + P2S</td>
<td>0.6957</td>
<td>0.7290</td>
<td></td>
</tr>
<tr>
<td>AAD + E2G</td>
<td>0.6935</td>
<td>0.7372</td>
<td></td>
</tr>
<tr>
<td>AAD + SS</td>
<td>0.6973</td>
<td>0.7969</td>
<td></td>
</tr>
<tr>
<td>AAD + SS + E2G</td>
<td>0.6995</td>
<td>0.8123</td>
<td></td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S</td>
<td>0.7018</td>
<td>0.8120</td>
<td></td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S + SN</td>
<td>0.7151</td>
<td>0.8128</td>
<td></td>
</tr>
<tr>
<td>AAD + SS + E2G + P2S + SN + G2E</td>
<td><strong>0.7166</strong></td>
<td><strong>0.8132</strong></td>
<td></td>
</tr>
</tbody>
</table>
4.4 Discussion

Our results demonstrated that our normalisation methods are effective in improving the performance of mapping concept mentions occurring in both clinical records and PubMed abstracts to appropriate entries in different terminological resources. Some examples of the correct mappings achieved are shown in Figure 4.14. The first term on each line is the textual concept mention, which does not correspond to any concept synonyms that are listed in either MeSH (light grey) or in SNOMED-CT (dark grey), while concepts in white are those occurring in both corpora, which do not correspond to concept synonyms in either resource. The original concept mention is followed by the successful normalisation technique that leads to a correct mapping, after which is the specific variant that is generated by this technique, and which allows mapping to the correct concept identifier in the relevant resource (as specified at the end of each line). As can be seen in Figure 4.14, for example, none of the textual concept mentions \textit{cardiac asystole}, \textit{elevated blood pressure}, \textit{reduced lung volume} and \textit{dystonias} corresponds to concept synonyms listed in MeSH. However, with the help of variants generated by various different methods (i.e., SS, E2G, SN and P2S respectively), all of these concept mentions can...
be successfully resolved to the correct MeSH IDs. Similarly, the concept mention *lung volume reduction*, which is a syntactic variant of the SNOMED-CT concept synonym *reduced lung volume*, can be successfully mapped to the correct CUI by using the SN method.

A further important point to note from the figure is that concept mentions that appear in both corpora (and hence which need to be mapped to appropriate concepts in both SNOMED-CT and MeSH) can be successfully mapped by our methods to the correct identifiers in both resources, which emphasises their flexibility. This is the case, for example, with *without strength, cholesterol elevation*, and *high WBC*.

As mentioned above, the only individual method which did not result in an improved F-score over the baseline (as reported in Table 4.9) was the P2S method, when applied to biomedical text. An error analysis revealed that although the majority of concept synonyms in the terminological resources correspond to singular forms, certain concepts that correspond to families or classes of more specific concepts are more properly specified using plural terms. Therefore, converting plural mentions to singular forms meant that they could no longer be matched. This was the case for *stress fractures* (D000662), *psychotic disorders* (D011618) and *progestogens* (D011372). Other concept mentions could not be mapped by our system to MeSH identifiers in either singular or plural form, according to the specific form of the concept synonym listed in the resource (e.g. *muscle contraction(s)*)), which is unlikely to match with many concept mentions in text. This could be addressed in future work through some preprocessing methods.

Although AAD achieved the highest F-score on texts from both domains, it was able to achieve a greater improvement over the baseline when applied to biomedical text (correctly mapping an additional 879 concept mentions, compared to the baseline), compared to clinical text. This is likely to be due to the more standardised use of abbreviations in biomedical text, along with the more formal nature of the text, and the fact that the AcroMine Disambiguator tool used was trained on biomedical abstracts. In contrast, clinical text contains a greater proportion of challenging ad-hoc abbreviations (which is particularly the case for the ShARe/CLEF corpus, as noted in Alnazzawi et al. (2016)), together with spelling errors and non-grammatical sentences, which can make disambiguation more challenging.
Figure 4.14: Examples of the improvement of automatic concept recognition in MeSH (light grey), the SNOMED-CT (dark grey), and both (white).
As we have already noted, all the hybrid methods outperformed the baseline for both corpora, as shown in Table 4.10. It is interesting to note that, although applying P2S individually to biomedical abstracts reduced the performance compared to the baseline, its combination with AAD could actually improve the F-score over AAD alone, with both increased precision and recall.

In both the clinical and biomedical text corpora, precision was not greatly affected by our methods, compared to the baseline. In the clinical corpus, the highest precision was achieved by the AAD+SS combination (0.7262 compared to the 0.7120 straightforward and 0.7151 MetaMap baselines). In contrast, in biomedical abstracts, even the baseline method produces very high precision (0.9727). Again, this stayed fairly stable, regardless of the normalisation method applied. Given the typically standardised terminology used in biomedical articles, most concept mentions that are matched exactly against concept synonyms in the terminological resource usually result in the correct mapping. In contrast, the more flexible and variable ways of expressing concepts in clinical text may result in greater ambiguity, and hence less precise mappings. Additionally, as noted in Alnazzawi et al. (2016), the gold standard assignment of CUIs in the ShARe/CLEF corpus often makes use of contextual information, rather than considering only the format of the mention itself, which can be problematic for our method.

For both clinical records and biomedical abstracts, recall increased consistently as more techniques were added into the hybrid method. In clinical records, the MetaMap baseline recall of 0.6766 was increased to 0.7184 in the complete hybrid method, while in biomedical abstracts, there was a more significant increase in the recall (from 0.7251 to 0.8358). This demonstrates the importance of considering a range of techniques to allow the normalisation of the greatest number of concept mentions. The greater increase in recall for the biomedical articles is likely to be due to the more limited and standardised forms of concept variants that appear in formal text, which are handled well by our systematic means of generating variants. In contrast, the highly variable and less predictable variations amongst concept mentions in clinical records, which frequently contain spelling errors (like amphotericin for amphothericin, and abscess for abscess), make the mapping process far more complex.

According to the nature of the language used in clinical records, the G2E method increases recall to a greater degree in this corpus than in biomedical articles.
However, the fine-grained nature of the concepts in SNOMED-CT and MeSH can sometimes mean that the Greek and English terms that are considered as equivalent by our methods (e.g., asthenia and fatigue) actually correspond to different concept entries, according to the subtle differences in their interpretation. The fact that the E2G method also has the greatest normalisation impact in clinical records suggests that concepts in clinical texts are described using a mixture of both formal and informal expressions. However, one reason for the fairly small contribution of this method towards overall normalisation performance is that a relatively small number of Greek variants was generated by the method. Specifically, for the 2,795 multi-word phenotypic concept mentions occurring in the clinical corpus, and the 1,748 in the biomedical corpus, only 591 and 398 Greek variants, respectively, were generated. This highlights the need to investigate further English and Greek element pairs and to add them to the medical suffix and prefix list.

Through error analysis, we also found that some of the gold-standard, expert-added normalisation annotations in both corpora were incorrect, meaning that the performance of our methods may actually be somewhat superior to that suggested by the results. In the BioCreative corpus, we found that some concept mentions were mapped to the wrong MeSH entries in the gold standard data. For example, dyspnea is the recognisable Greek variant of the original concept mention difficult breathing, which is not listed in MeSH. However, the unique ID of the correct concept (D004417: dyspnea) does not match the gold standard ID, which actually corresponds to a more general concept (D012120: respiration disorders). In contrast, in the gold standard data for clinical records, a large number of concept mentions were marked by the expert annotators as “CUI-less”, i.e., they cannot be mapped to existing concepts in SNOMED-CT. However, we found that many of these concept mentions are actually assigned a CUI in SNOMED-CT by our methods, which appears to be correct. For example, our methods are able to assign CUIs to the following terms, which are annotated as CUI-less in the ShARe/CLEF corpus (and hence these assignments are counted as false positives in the evaluation of our method): aphagia (C0221470), brain death (C0006110), choking (C0008301), instability (C1444783), ecchymosis (C0013491).

As has been mentioned above, both of the corpora used in our evaluation were developed in the context of shared tasks, and have been used to evaluate the performance of several other normalisation methods. However, while the evaluation
of our methods has concerned measuring normalisation performance on pre-identified concept mentions, the original shared task evaluation results were based on a more complex task, i.e., the combined task of both the automatic recognition of concept mentions and their normalisation. This means that our results are not directly comparable with those of the other methods that have been applied to the same corpus. For reference, however, here we considered MetaMap as the baseline and applied it to the same corpus that was used to develop our hybrid method. As shown in Table 4.12, higher F-score (0.89) and accuracy (0.81) are achieved by our methods. Besides, the highest performing method in the BioCreative V Chemical Disease Relation (CDR) Task was based on machine learning (a CRF model with post-processing) and achieved an F-score of 0.865 (Wei et al., 2016). Given that machine-learning methods generally perform better than other methods, we believe that the performance of our non-machine-learning based hybrid normalisation method (0.891 F-score) is highly respectable. As a further point of comparison with the performance of our method on biomedical text, the PhenoNorm normalisation method (Alnazzawi et al. 2016) was evaluated on the NCBI disease corpus (Doğan et al. 2014), which, similarly to the BioCreative corpus, consists of normalised disease mentions in biomedical abstracts. On this corpus, while the PhenoNorm method achieved 0.64 accuracy (0.69 F-Score) in normalising pre-recognised concept mentions, our hybrid method reached an accuracy of 0.81 on the BioCreative corpus (as shown in Table 4.12). The lower performance compared to that of our hybrid method on the BioCreative corpus is an evidence that our combination of multiple normalisation techniques is better able to handle the various types of fairly systematic variations amongst concept mentions that are typical in biomedical text.

Table 4.5: Comparison of our hybrid method against other approaches applied to the biomedical corpus.

<table>
<thead>
<tr>
<th>Normalisation method</th>
<th>Corpus</th>
<th>F-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaMap</td>
<td>BioCreative corpus</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>PhenoNorm</td>
<td>NCBI disease corpus</td>
<td>0.69</td>
<td>0.64</td>
</tr>
<tr>
<td>Wei et al., 2016</td>
<td>BioCreative corpus</td>
<td>0.87</td>
<td>-</td>
</tr>
<tr>
<td>Our hybrid method</td>
<td>BioCreative corpus</td>
<td><strong>0.89</strong></td>
<td><strong>0.81</strong></td>
</tr>
</tbody>
</table>
In the ShARe/CLEF task, overall results were calculated based on accuracy, and in the original task, the best performing system achieved an accuracy of 0.589 in recognising and normalising disorder concept mentions; this fairly low score reflects the challenging nature of processing this text type. However, in the task of normalising pre-recognised concept mentions, PhenoNorm achieved an accuracy of 0.83 compared to 0.72 for our best hybrid method. The comparison is shown in Table 4.13. An interesting contrast between PhenoNorm and our approach is that, whilst our method has the greatest impact on normalisation in biomedical text (compared to the MetaMap), PhenoNorm has a greater impact on normalisation in clinical text. This is likely to be due to the use of string similarity methods in PhenoNorm, which allow, for example, flexible mapping of the concept mention *stenosis in left anterior descending* to the synonym *left anterior descending coronary artery stenosis*, and which allow mapping of concept mentions containing spelling errors. However, as we have previously mentioned, the restriction of matches to semantically equivalent phrases, which is a feature of our method, is also highly important, and can help to prevent some of the mapping errors made by PhenoNorm. Furthermore, the MetaMap method is applied to the same corpus as used in our project (i.e., the ShARe/CLEF corpus), and our method increased F-score and accuracy from 0.70 and 0.64 to 0.72 and 0.72, respectively.

Table 4.6: Comparison of our hybrid method against other approaches applied to the clinical corpus.

<table>
<thead>
<tr>
<th>Normalisation method</th>
<th>Corpus</th>
<th>F-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaMap</td>
<td>ShARe/CLEF corpus</td>
<td>0.70</td>
<td>0.64</td>
</tr>
<tr>
<td>Suominen et al. 2013</td>
<td>ShARe/CLEF corpus</td>
<td>-</td>
<td>0.59</td>
</tr>
<tr>
<td>PhenoNorm</td>
<td>-</td>
<td>-</td>
<td>0.83</td>
</tr>
<tr>
<td>Our hybrid method</td>
<td>ShARe/CLEF corpus</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>
4.5 Summary

In this chapter, we have described our development of a number of different concept normalisation techniques that account for a range of variations that can occur amongst concept mentions in texts belonging to different domains. Our results have demonstrated that the methods are robust for application to both clinical and biomedical text, in that they are able to achieve a significant improvement in normalisation performance over a baseline system applied to corpora of both text types. By appropriately combining the techniques into hybrid method, we were able to achieve an improvement in normalisation performance over MetaMap baseline system applied to corpora of both text types. The resulting hybrid normalisation method achieved maximum F-scores of 0.7208 for clinical text and 0.8908 for biomedical text. These scores represent increases of 3.67% and 10.19%, respectively, over the MetaMap baseline approach.
Chapter 5

Conclusion

In this chapter, we revisit the research objectives that have been stated in Section 1.2 and briefly explain how we accomplished each of them. In addition, we also share our plans and suggestions for extending our work in the future.

5.1 Evaluation of research objectives

To gain the state-of-the-art knowledge in phenotypic named entity recognition, we established the objective $O_1$:

$O_1$ To conduct a comprehensive review of existing annotated corpora and annotation approaches for phenotypic named entity recognition (NER).

to address the research question $Q_1$:

$Q_1$ What are the existing approaches for phenotype extraction and annotation?
As an initial step towards achieving this objective, we analysed previously reported pre-processing (Section 2.1.1) and post-processing approaches (Section 2.1.2) to NER, and selected truecasing and abbreviation disambiguation as our pre-processing methods, and distributional thesaurus lookup as the method for post-processing. Since a combined annotation system was demonstrated to considerably outperform any of the individual systems, hybrid combinations of these three methods using two refinement schemes (i.e., sequential and parallel) were then investigated.

By reviewing the named entity recognisers used in biomedical and clinical domains in Section 2.1.3 and 2.1.4, separately, we determined that machine learning methods, such as the CRF algorithm has become one of the most commonly used approaches for NER. This enabled us to select NERsuite, which is based on the implementation of CRFs, as our NER tagger. Available biomedical and clinical corpora known to us were also analysed in Section 2.1.5, and finally we chose the 2010 i2b2/VA challenge data set in our research (as described in Section 2.2).

These results supported us in fulfilling our second objective $O_2$:

$O_2$ To develop phenotypic named entity recognisers to extract phenotypes in biomedical and clinical articles.

In Section 2.3, we described our machine learning-based phenotypic named entity recogniser. Unlike most concept extraction systems (as described in Section 2.1.6), we did not use any external knowledge resources since annotation is always a time- and labour-consuming task, and the public could not straightforwardly access most of those corpora.

Implementing this idea, we incorporated three pre- and post-processing methods and developed our own phenotypic named entity recogniser. While truecasing and abbreviation disambiguation capture the morphology of words, the distributional thesaurus lookup allows for statistics-based similarity matching. We achieved a maximum improvement of F-score from 0.7565 and 0.8408 to 0.7586 and 0.8444 for exact and relaxed matching, respectively. The obtained improvements are modest as our preprocessing or post-processing components are added to the same core information extraction system (i.e., NERsuite), but the results still indicate that
truecasing and annotation combination are the enhancements which best increase the system performance whilst the transformation of short-form terms to their suitable long-form candidates can reduce the word ambiguity and improve the performance as well. These results are reported in the Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (Fu and Ananiadou 2014). We have hence answered the second research question Q₂:

**Q₂** What improvements can we introduce to the current methods to achieve better performance for phenotype extraction?

Based on those results, we established the following objective O₃ as a step towards our study on biomedical and clinical phenotype annotation.

**O₃** To develop a phenotype annotation workflow with the integration of various text mining techniques.

By reviewing previously reported corpus annotation approaches in Section 3.1, we addressed the third research question Q₃ we have raised:

**Q₃** What are the state-of-the art approaches for medical corpus annotation?

Since COPD is a life-threatening lung disorder whose recent prevalence has led to an increasing burden on public healthcare and its phenotypic information is essential in providing suitable personalised treatment, we decided to develop two corpora that contain clinical and biomedical articles with annotation of COPD phenotypes, respectively. We accomplished the third objective by dividing the problem into two subtasks, namely, document collection (Section 3.2) and annotation (Section 3.3).

The clinical corpus consists of clinical records of 50 patients with COPD collected from a British hospital (Section 3.2.1); while the corpus of biomedical literature is a collection of 30 full-text articles from the PubMed Central (PMC) Open Access subset (Section 3.2.2).

In order to support the process of COPD phenotype curation, we described our annotation scheme for capturing COPD phenotypes, which is aimed at producing
fine-grained, expressive and computable COPD annotations without burdening our curators with a highly complicated task in Section 3.3.1, and in Section 3.3.2, we presented a semi-automatic annotation workflow that integrates several text mining tools, including the truecasing module which was developed and proved to achieve the best performance of concept extraction in Chapter 2. The outcome of these two corpora development efforts was reported in the Proceedings of the Phenotype day at ISMB 2014 (Fu et al. 2014) and the Journal of Biomedical Semantics (Fu et al. 2015), respectively.

Normalisation is essential for linking phenotypes in clinical records of individual patient with biomedical knowledge bases, enabling clinicians to provide personalised treatments. In order to explore phenotype normalisation as a response to the fourth research question Q₄:

\[Q₄\]  What are the differences between phenotypes in biomedical and clinical domains?

In the beginning of this research, we also established the objective O₄:

\[O₄\]  To analyse the differences between the phenotypes extracted from biomedical and clinical domains.

This objective was accomplished in Chapter 4 as clinical records consist primarily of brief narrative descriptions, acronyms and abbreviations, ungrammaticality, and misspelling while biomedical literature is more academic and technical, mostly expressed in professional medical sublanguage derived from Greek language.

The following objective O₅ was established:

\[O₅\]  To conduct a comprehensive review of existing techniques for concept normalisation.
to answer our fifth research question $Q_5$:

$Q_5$ Which normalisation approaches can be applied to those two heterogeneous domains?

A review of existing approaches to concept normalisation was completed and presented in Section 4.1. Due to the availability of annotated corpora, all those methods were evaluated in a single domain, and their authors were not able to investigate the differences or the relationships between those two domains, i.e., the clinical and the biomedical domains. Our research, therefore, is the first work reported on the normalisation between clinical records and biomedical literature with the end-goal of providing tailored treatments for individual patients.

Since there is a significant variation in expressions comprising concepts from both clinical and biomedical domains, we have developed a novel, flexible method, which is able to normalise concept mentions with a wide range of characteristics in heterogeneous text types. We have taken particular care to allow links to be made between the Greek-derived terms that are common in biomedical articles and the more descriptive English phrases that are characteristic of narrative clinical text. We also address the ambiguous nature of acronyms/abbreviations, which can be particularly prevalent in clinical text. Four different types of variation generation methods have been developed (i.e., compositional synonyms (Section 4.2.1), normalisation of syntactic structure (Section 4.2.2), acronyms and abbreviations (Section 4.2.3) and plural and singular (Section 4.2.4)), which was stated in our sixth objective $O_6$:

$O_6$ Based on the analysis fulfilled by $O_4$ and $O_5$, to develop implementations of different normalisation methods and observe their impact on linking performance.

Addressing the following research question $Q_6$, we demonstrated that the resulting hybrid normalisation method performs robustly when applied to both narrative clinical text and biomedical academic text, resulting in a significant improvement in normalisation performance. As described in Section 4.3, the F-score
was increased by 3.67% and 10.19% from the MetaMap approach for the clinical and the biomedical domain, respectively.

Q6 Does the normalisation improve the performance of linking clinical records with biomedical literature?

After having fulfilled our research objectives, we obtained the findings summarised above which prove the two research hypotheses H$_1$ and H$_2$ that we proposed at the beginning:

H$_1$ Phenotypes in both biomedical and clinical domains can be recognised using existing concept extraction methods and the performance can be improved;

H$_2$ Even though phenotypes in biomedical and clinical documents are distinct, they can be normalised and identified.

5.2 Future work

Given the promising nature and demonstrable performance improvements achieved by our methods that generate English equivalents of Greek terms, and vice versa, part of our future work will involve the development of a more complete list of medical prefixes, suffixes and their language roots, such as the lexical databases distributed with the UMLS SPECIALIST Lexicon. This will allow a greater number of potential Greek and English variants of concept mentions to be generated, which should ultimately lead to an improvement in normalisation performance. Although several lists of Greek and Latin roots have been described (Green 2014), they are not specialised for the clinical or the biomedical domains. Thus, the meanings of each root need to be further clarified.

Additionally, given the demonstration that string similarity methods can be helpful in dealing with the highly variable nature of concept mentions in clinical text, along with spelling errors, we will also consider how these can be successfully integrated within our method, whilst ensuring that semantic equivalence amongst generated variants is maintained.
Finally, we intend to demonstrate how normalisation can be used to facilitate the efficient linking and integration of complementary information from biomedical abstracts and clinical records, initially by making use of 50 clinical records of COPD patients that we have collected from our collaborating hospital.
References


REFERENCES


ATS 1995. Supplement on Standards for the Diagnosis and Care of Patients with Chronic Obstructive Pulmonary Disease. American Journal of Respiratory and Critical Care Medicine 152: s77–s120.


REFERENCES


REFERENCES


REFERENCES


Appendix A

List of COPD phenotypes used to retrieve articles from the PubMed OpenAccess subset

postoperative pulmonary rehabilitation, global initiative for chronic obstructive lung disease, advance care decision, visible nasal cannula, lung syndrome, maintenance antibiotic therapy, co-oximeter, population average, energy requirement, pulmonary function, arterial oxygen tension/inspiratory oxygen fraction, assess risk, o2 and transcutaneouspa, lung volume reduction surgery, hypoxic vasoconstriction, concentrator oxygen system, flow requirement, pulmonary muscular artery, moderate-to-severe copd patient, body weight, nasal mucosa, neutrophilic inflammation, anesthesiology physical status scale, exercise endurance, complication rate, home oxygen, cardiovascular instability, optimal medical regimen, bilateral lung volume reduction surgery, frequent arousal, eosinophilic inflammation, intermediate intensive care unit, pneumococcal disease reduces bacteraemia, simple timed walking, symptoms and bacteriological resolution, large portable oxygen system, emphasis shift, pulse oximetry, nutrition examination survey, NO, infrequent exacerbator, decision analysis, regular ics.alway, time-limited trial,
permanent reduction, computed tomography scanning, behavioural change, dynamic activity, airways disease, perceptual motor, impaired skeletal muscle strength, bilateral surgery, peripheral airway collapse, quit plan, 90-day surgical mortality, near-maximal exercise, short-acting bronchodilator, pharmacological support, abdominal surgery, common indication, nonresected area, long-term oxygen therapy outweigh, nasal irritation, severe acute hypoxaemia, superoxide anion, OSA, radiographical high-resolution, ambulatory oxygen therapy, maximum oxygen consumption, pulse demand, rapid eye movement, transtracheal oxygen delivery, sleep abnormality, muscle activity, rehabilitation patient, high-intensity activity, postoperative humidification, medical regimen, healthcare utilisation, sleep-related hypoventilation, moderate exacerbation, patient population, social performance, emergency room therapy, cell infiltrate, oxygen-induced hypercapnia, oxygen saturation, renal and liver failure, acute care, low dyspnoea, 4-mgnicotine dose, full medical regimen, tobacco-use status, negative pressure ventilation, anaesthetic agent, preventative strategy, smoking cessation, gradual tapering, pulmonary mechanic, first-line intervention, sleep-related hypoxaemia, exacerbation of copd, plateau phase, sea-level pa, anabolic steroid, oxygen setting, alveolar hypoxia, adequate oxygenation, a central bronchus, crq-measured health status, usual exertion, standard didactic session, smoking addiction, oxyhaemoglobin dissociation curve, severity of copd, physicaland totalsip, flow oxygen, severe disease, nocturnal arterial oxygen saturation, poor washout, copd progression, reversible airflow limitation, postbronchodilator fev1, nicotine base, "alveolar attachments."" OR ""tracheal ring", normal pa, passive smoke exposure in childhood, outpatient and home-based rehabilitation, family history of respiratory disease, respiratory-specific health status questionnaire, smooth muscle contraction, tissue inhibitor, lean body mass anxiety, alveolar walls, medication fev1lung, bullous emphysema, treatment response criterion, small epithelial gland, moderate-to-severe copd, absorptive at melectasis, exercise training, oxygen-conserving device function, care planning derive, newer pulse oximeter, ex-smoker, active treatment, first-line options, arterialoxygen tension, pulmonary vasodilator therapy, bronchial reactivity, oral calciumantagonist, mucolytic/antioxidant therapy, airways obstruction, antioxidant defence, terminal condition, long-acting neuromuscular blocker, hospital anxiety and depression questionnaire, staph, low zone, smoker, loss of muscle mass, moraxella catarrhalis, cardiac arrhythmia, diaphragmatic contraction, ats statement, gas volume,
bronchodilator drug, BODE index, ventilatory support, subglottic stenosis, motor skill, standard nasal cannula, specialist care, pulmonary arteriole, HADS, oral steroid, airway disease, arterial hypoxaemia, lb reservoir, nasal cannula delivery, cancer patient, end-of-life decision-making, interval exercise training, progressive respiratory acidosis, respiratory illness, goblet cells, hypoxemic patient, long-term adherence, cell count, health-related quality, FVC ratio, postoperative recovery period, tobacco user, airflow limitation, exposure history, oxygenate patient, biological molecule, bythe tdi, obliterative bronchiolitis, rich body, chronic co2 retainer, cor-pulmonale, gene-environment interaction, standard medical therapy, psychosocial status, external pressure support, passive tobacco smoke, initial pa, morbid obesity, sleep fragmentation, explosive disconnection, dark skin pigment, metabolic abnormality, abolish symptom, neuroaxial blockade, hypoxaemic patient, recheck abg, pulmonary artery, anabolic stimulus, high-dependency unit, squamous metaplasia of the airway epithelium, combination therapy, antibiotic therapy, inhalation device, poor prognosis, secondary complication, dyspnoea management, co2 and acid base status, anticholinergic drugs, chronic oxygen, ambulatory copd patient, evidence-based approach, follow-up intervention, medical research council dyspnoea scale, end-of-life support, preoperative blood gas, lung cancer, subjective complaint, exercise tolerance, relative contraindication age, involuntary weight loss, o2in patient, rehabilitation group, financial burden, oxygen-conserving system, teenage smoker, socioeconomic background, environmental risk, patient surrogate, congestive cardiac failure, preoperative pulmonary function, social burden, outpatient therapy, chronic productive cough, diaphragm muscle mass, inhaled corticosteroids, behavioural intervention, accessory respiratory muscle, cor Pulmonale, nicotine replacement therapy, exacerbation episode, acute hypobaric hypoxia, anti-oxidant action, permanent abstinence, long acting muscarinic agonist, small minority, aggressive medical treatment, recommended vaccine, uncomplicated copd, cardiovascular collapse, prominent intimal thickening, muscle deconditioning, cost/utility analysis, gold standard method, nonspecific symptom, cerebrovascular disease, air pollution exposure, protein balance, double lung transplant, dioxide tension, muscle wasting, airway remodelling, peripheral vasculature, acute hospitalisation, severe gas, nocturnal supplemental oxygen, bacterial, oxygen toxicity, concentrator, sleep desaturation, mild hypoxaemia, 100-point scale, orthopedic procedure, oxygen desaturation, physiological hallmark, physiological
adaptation, cessation activity, multidisciplinary team, stress management, psychopathological impairment, complex pharmacokinetic, predominant emphysema, muscle fatigue, age-specific rate, simple screening, the catheter, ambulatory oxygen system, perhydroxy radical, low water vapour output, haemodynamic instability, palliative care resource, arterial blood, mucus ball formation, chronic sinusitis, carbon dioxide tension, spacer chamber, low oral ph, oxygen flow rate, carbon monoxide, monoamine oxidase inhibitor, catheter blockage, small bolus, copd severity, nicotine replacement, impact component, atmospheric gas, exercise performance, sustained pulmonary hypertension, nonspecific phosphodiesterase inhibitor, overnight sleep, healthcare professional, flow setting, symptoms in patient, hospice care, close relationship, first-line treatment, forced vital capacity, time-limited span, depressive symptoms, ventilatory effort, ischemic heart disease, pre-flight testing, host factor exposure, full humidity, a change in their sputum characteristic, international carrier, bone mineral density, daily energy expenditure, polacrilex resin, diagnosis suggestive feature, modest dose-response, domiciliary care, unexplained weight loss, bronchial gland hypertrophy, cardiac function, most programme, arterial oxygen, ventilation-perfusion inequality, 24-h formulation, integral component, pressure support, hypercapnic respiratory failure, questionnaire component, non-exacerbator, postoperative respiratory insufficiency, simple spirometry, streptococcus pneumoniae, transdermal nicotine systemor, alternative delivery device, v moribund, high-flow transtracheal catheter, empirical guidance, nursing facility, healthy lifestyle, purulent sputum, sustained lung function, hypoxaemic range, concurrent psychiatric morbidity, modular document, dynamic hyperinflation, laser treatment, somenasal airflow, oximeter display, medical management, obesity hypoventilation, muscle contractility, alveolar space, nasal/oral dryness, gas system, electronic medical record, regular exercise, postrehabilitation 6-min walk, respiratory symptom, treadmill ergometer, decision-making process, viathe telephone, abdominal vascular surgery, oxygen consumption, actual fi, bupropion preparation, severe lung disease, bronchial biopsy, physical activity, venous thromboembolism, cor pulmonale, anticholinergic agent, high-risk comorbid condition, lung hyperinflation, family training, cigarette smoking, relapse prevention, activity pattern, E Coli, moderate-to-advanced copd patient, relative with qualifier, aversive training method, gas transfill portable withocd, anticonvulsant drug, genetic predisposition, upper abdomenare, oxygen therapy, evidence-based guideline,
exercise prescription, liver cirrhosis, benefit from palliative service, minute ventilation, bronchodilator therapy, metallic taste, status using the sgrq, lvrs patient, steep portion, exchange abnormality, breath alarm, nocturnal oxygen therapy, low post rehabilitation exercise capacity, tobacco smoking, transfill system, end-of-life education, oximetry sp, lung function via ventilation, terminal sedation, arterial carbon dioxide tension, Requirement for invasive mechanical ventilation, spacer device, excessive breathlessness persist, hydrogen peroxide, sputumour dyspnea, terminal bronchiole, patient-centered palliative care, caloric load, acid-base information, metered-dose inhaler, matrix metalloproteinases, facial skin erythema, carbon dioxide, heat blister, extreme obesity, expiratory volume, severity of exacerbation, hypoxic origin, strength influence exercise capacity, co-existing copd, encouraging discussions thin family, eosinophilic, modest yearly decline, high-dose steroid, moderate copd, perioperative period, disease risk, protein synthesis, international consensus statement, nicotine replacement treatment, catheter displacement, genetic risk factor, component of pulmonary rehabilitation, massive pleural effusion, psychological change of chronic illness, high-flow enriched oxygen, dyspnoea via a reduction, noninvasive mechanical ventilation, home caregiver, chronic disease management, baseline dyspnoea, end-organ dysfunction, gold initiative, comorbid problem, arterial carbon dioxide tension, exposure to hypobaric hypoxia, tricuspid insufficiency, bronchitis, intermittent positive pressure ventilation, experience neuropsychiatric deficit, physical activity can, pharmacological therapy, peptic ulcer, systemic steroid, deep vein thrombosis, diabetes mellitus, invasive mechanical ventilation, composite picture, mosaic pattern, adrenal insufficiency, submaximal testing, improved dyspnea, postrehabilitation baseline score, pulmonary embolism, type II respiratory failure, continuous oxygen, ers standard, stable copd patient, abundant deposition, renal system, reliable screening technique, obligation of insurer, nasal oxygen delivery, airway hyper responsiveness, nocturnal oxygen desaturation, panic control, epithelial cell, accurate diagnosis, mucous secretion, emphysematous patient, chronic allograft rejection, continuous assessment, bmi, environmental condition, strenuous exercise, regression approach, nicotine dependence, c-reactive protein, normal oxygen metabolism is water, respiratory failure, hepatitis b antigen positivity, shortness of breath, doppler echocardiography, o2 flow rate, co-morbid condition, therapeutic response, lung volume-reduction surgery, neutrophil elastase, co-morbid disease condition, smoking initiation,
preoperative evaluation, human service guideline, low respiratory tract colonisation, respiratory depression, heart-lung transplant, mucous gland hyperplasia, blood perfusion, life-supportive care, marked effectiveness, ventilatory failure, oxygen delivery method, evaluable died, portable compressed gas oxygen cylinder, intravenous anaesthetic, dependence guideline, repeat complementary exam, chest/abdominal wall, inhaled steroids, respiratory muscle strength, behavioural support, maximal-achieved cycle ergometry, obese patient, added benefit, common comorbid, dietary intake, clinical rapid progressive dyspnoea, pulmonary circulation, fire-retardant plastic, assurance of terminal sedation, unos on-line data base, COPD assessment test, liquid oxygen, mild-to-moderate copd, nicotine dose, oral corticosteroid, respiratory sign, pulmonary vascular pressure, closed angle glaucoma, refractory hypoxaemia, tobacco dependence, progress neutrophil, respiratory centre, physiological assessment, pharmacological treatment, exercise testing, movement artifact, hypoxia inhalation test, comprehensive intervention, laboratory test, buccal mucosa, angina, once-daily dose, cell blood count, attentive clinician, first-line pharmacotherapy, breath-activated inhaler, long-term oxygen rx, monoclonal antibody, postoperative pulmonary complication, noxious agent, premature mortality, upper lobe predominant emphysema, physical suffering, inpatient care, respiratory rate, undiagnosed copd, quit attempt, combination LABA and ICS, abnormal gas exchange, noninvasive positive pressure ventilation, peripheral muscle wasting, stroke, hand-held nebuliser, brief intervention, tracheal obstruction, beta-blocker medication, cause of death, ventilation-perfusion relationship in exacerbation, supportive end-of-life care, irreversible disease, tolerate co2 retention, inflammatory response, performance improvement effort, disease onset, continuous oxygen therapy, antibiotic treatment, preoperative chest radiography, non-upper lobe, energy balance, bacterial infection, alveolar/arterial o2 ratio, mononuclear cell, increased resistance of the small airways, mouth breathing, provide material, noninvasive method, energy-dense supplement, clinical factor, life closure, self report, direct healthcare cost, gastric irritation, genetic factor, permanent enlargement, endothelial dysfunction, head/neck procedure, pregnant woman, malhospice service, drug treatment, severe copd experience, post-rehabilitation cycle ergometry, economic burden, second-line treatment, subcutaneous emphysema, plasma cell, isoprostane, decreased dl, sustained quitter, inflammatory cell, mucus hypersecretion, cycle ergometry, operational classification, nonphysician rehabilitation caregiver,
paradoxical bronchoconstriction, commercial airliner, oxygen litre flow, cachexia, hypobaric exposure, craniofacial trauma, co-existing sleep apnoea syndrome, pulsed demand nasal delivery, transtracheal oxygen, weight change, abnormal diffusing capacity, airway pressure, noncartilaginous airway, high-flow nasal oxygen, GOLD, combined application, pulmonary function test, terminal lung, oxygen delivery, reservoir cannulae, co2 and ph, mortality database, permissive hypercapnia, tobacco abuse, severe dyspnoea, nocturnal hypoxaemia, lung volume, pulmonary complication, nocturnal cardiac dysrhythmia, haemodynamic, unoxygenated blood, respiratory disease, progressive chronic disease, nocturnal oxygen therapy trial, copd exacerbation, patient education, pulmonary vasculature, cell dysfunction, oxygen-conserving device, phase inhalation, copd death rate, equicaloric fat-rich supplement, peripheral muscle, chronic hypoxaemia, protein turnover, pulmonary gas exchange, eye irritation, standard transfuse system, smoking, cortical input, unstable angina, cabin environment, brief tobacco dependence intervention, home care, submaximal workload, WCC, exacerbation frequency, exacerbation section, palliative intervention, survival advantage, respiratory system, high-dose inhaled corticosteroid, potential mechanisms, sleep disturbance, institutional support, streptococcus pneumonae, palliative care, clinical practice, inadequate preparation, acute respiratory failure, ventilation/perfusion mismatch, permanent remission, asthma early onset, patientable to resume ambulation, tissue perfusion, respiratory acidosis, wedge resection, lung volume change, copd outpatient, nocturnal sa, intensivist or pulmonologist, close monitoring, movement artefact create inaccuracy, ventilation-perfusion matching, maximum in inspiratory pressure, normal physical examination, clinician diagnosis lack sensitivity, health-related quality of life, iron lung, pulmonary capillary bed, alveolar volume, renal or liver failure, connective tissue, myocardial infarction, insertion site, respiratory therapist, primary disorder, patient-family-friend communication, actual dose may vary, lung compliance, loss of appetite, transcutaneous arterial oxygen tension, inspiratory oxygen, symptoms of depression, occupational hazard, nonrebreather mask, life supportive intervention, GORD, lung disease, superior analgesia, basic source document, tobacco use, strength training, low pretreatment, pharyngeal deposition, end-of-life discussion, typical fi, constant fi, short interval, upper abdominal surgery, severe acidosis, rehabilitation service, computer prompt, ventricular function, orthopaedic procedure, high-risk procedure, carbon monoxide per litre, maximal exercise, chronic
obstructive pulmonary disease, tobacco intervention, long acting beta agonist, multiple ill-defined bullae, transitional dyspnoea index, evidence based guideline, local tissue perfusion, mild daytime hypoxemia, chronic condition, active participation, 6-min walk distance, quality-of-life benefit, elevated energy metabolism, chronic lung disease, structural change, peak oxygen consumption, mastery component, training duration, dose inhaler, obstructive sleep apnoea, mechanisms of hypoventilation, quality meta-analysis, bilateral lvrs, practical purpose, thick bundle, clear responsibility, rate of lung function decrease, improved sgrq, CXR, perfusion scintigraphy, emergency room discharge, exacerbation severity, worldwide contribution, exacerbation, topical side-effect, local irritation, arterial o2 saturation, previous smoker, sex socioeconomic status, harmful tobacco product, histological section, high-dose nebulisation, seizure disorder, squamous metaplasia, copd desire assistance, under-ventilated alveolus, o2 during exertion, inpatient oxygen therapy, upper extremity training, smoking patient, extracellular matrix, accessory muscle, cardiovascular status, home oxygen therapy, gastroesophageal reflux disease, pulmonary rehabilitation programme, respiratory bronchiole, exercise limitation, alveolar pressure, a normal ph, fast decline, breath rate, several controlled trial, diagnostic centre, staphylococcus aureus, coronary vascular disease, peripheral airway, severe oxygen desaturation, supplemental oxygen, simple face mask, conventional end-point, characteristic pathological lesion, smoking habit, nutritional intervention, dry mouth, mucous gland, prostatic symptoms, awake arterial oxygen, smoking history, smooth muscle cell, initial oxygen delivery, primary care physician, exposure to pollutant, healthcare service, pollution, rem-associated fall, arterial o2 tension, airway smooth muscle, surgical mortality, anticholinergics, sputum production, acid-base balance, perioperative pulmonary dysfunction, exhalation early inspiration delivery store, co2 retention concern, aggressive respiratory care, postoperative pneumonia, reasonable expectation, absolute contraindication, productive chronic cough, right heart failure, pulmonary rehabilitation intervention fall, bronchial gland, ave sainte-luce, measure pa, FVC, nocturnal respiratory failure, residual capacity, lifes upporte care, particulate burden, alveolar hypoventilation, severe bronchitis, light ambulatory system, respiratory condition, smoking-related emphysema, body temperature, CRP, unilateral lvrs, improved pulmonary function, borderline pa, gas oxygen container, general consensus, ventilation-perfusion ratio, low intensity exercise training,
arterial carbon, cabin altitude, long-term oxygen therapy, residual volume, family screening, lung capacity, nonbronchodilator activity, gastro-oesophageal surgery, "sa, o2", cardiovascular screening test, passive smoking, high-flow oxygen, acute response, emergency department visit, withdrawal symptom, thoracic hyperinflation, pharmacological regimen, rehabilitation recommendation, energy intake, sole determinant, venous blood, long flight, stationary oxygen, comorbid condition, nocturnal pulse oximetry, femoral neck mineralisation, monitor patient, current smoker, quit rate, durable power, decubitus ulcer, low postrehabilitation, sleep apnoea, hypercarbia concern, air travel, alveolar support, progressive copd, exercise ability, substance addiction, end-stage disease management, observational cohort, nicotine nasal spray, postoperative management, treating tobacco use, delivery efficacy, chronic respiratory questionnaire, oxygen requirement, occupational therapist, tobacco-dependence treatment, mucous hypersecretion, operative risk, oxygen tension, post-bronchodilator FEV1/forced vital capacity, desirable element of pre-flight patient care, copd patient, airway calibre, pulmonary fibrosis, impaired balance, lung tissue, unpressurised cabin, shorter-acting atracurium or vecuronium, diminished responsiveness, high flow via a transtracheal catheter, regulatory agency, pulmonary disease, respiratory tract infection, cessation intervention, dual energy x-ray absorptiometry, candidate gene, ciliary dysfunction, capillary blood, lVRS medical, essential component, white female, respiratory arrest, infectious process, walk distance, long history of smoking, obesity, dose therapy, airway reactivity, ICU rank pain control, oxygen supplementation, initial flow, face mask, smooth muscle relaxation, biomass fuels, distal airspace, heterogeneous emphysema, quality improvement, LTOT indication, holistic approach, haemodynamic stability, nonupper lobe, respiratory questionnaire, nutritional supplement, microscopic lesion, independent predictor, airway obstruction, tobacco product, standardised method, end-of-life planning, particular attention, reimbursement criterion, flat dose-response relationship, fat mass, preferred measure, subsequent local side-effect, ventricular dysfunction, exertion and sleep, dry powder inhaler, type 2 respiratory failure, transfusion requirement, far-advanced disease, air carrier, definitive recommendation, nett research group experience, lung transplantation, sleep benefit, progressive muscle relaxation, surgical complication, arterial oxygen saturation, nutritional assessment, noninvasive sp, small airway dysfunction, b celli, long term oxygen therapy, narrow therapeutic margin, nutritional status, second-line options
for oxygen delivery, obvious fibrosis, clinical evaluation, intensity training, day-to-
day variability, chest wall surgery, polysomnographic sleep, tobacco control, pleural
effusion, intratreatment social support, effective and, ocular pressure, free radical,
outpatient treatment, several strategy, clinical suspicion, respiratory medication,
management of stable copd, low-flow oxygen, occupational exposure, pulse demand
device, hospital discharge, sleep oxygen, difficulty sustaining, cardiovascular
surgery, surgical procedure, lifetime commitment, moderate acidosis, airway wall,
hyperinflation, pulmonary nodule, inflammatory change, occupational pollutant,
hallmark symptoms, placebo medication, nutritional therapy, extreme arterial blood
gas abnormality, early mobilisation, weak cough, nutritional supplementation,
consecutive enrollment, oxygen titration, fat-free mass, bronchodilator response,
haemophilus influenzae, asthma and copd, physiological evaluation, differential
diagnosis, cd8+ type, quit attemptor, bullae, impaired health status, needle stick,
chronic inflammation, chronic bronchitis, specific schedule, energy expenditure,
skeletal muscle myopathy, pathological extension, acid base status, asthma
occupation, capillary bed, effective management, respiratory dysfunction, portable
oxygen, storage capacity, blood gas testing, frequent exacerbator, physical hazard,
sip score, short burst oxygen therapy, ophthalmologic procedure, modern slow-
release preparation, low extremity training, postoperative mechanical ventilation,
upper airway, indistinguishable from patient, relative contraindication, chronic sleep
disturbance, post rehabilitation cycle ergometry, chronic rejection, outpatient
rehabilitation, prevalence of nutritional depletion, rem sleep, pressure support
ventilation, clinical efficacy, forced expiratory volume, comprehensive programme,
dose-response relationship, oxygen uptake, severe copd, primary end-point,
household member, heath status, nicotine polacrilex, endotracheal intubation, end-
stage copd, young patient, significant minor toxicity, ATS, panel consensus,
pulmonary hypertension in copd, inspiratory oxygen fraction, fear of failure,
premature ventricular contraction, co-existing sleep apnoea, spirometric
improvement, specific oral peripheral chemoreceptor stimulant, concurrent alcohol,
prepare patient, stress reduction, abnormal distribution, simultaneous resection,
transdermal nicotine, lung function testing, nutritional support, specific section,
augment co2 removal, skeletal muscle dysfunction, chronic care, noxious particle,
adequate daytime oxygenation, evaluate improvement in symptoms and physical
exam, noninvasive ventilation, comprehensive advance care planning, emotional
disturbance, surgical constraint, thoracic epidural anaesthesia, pleural pressure-time index, cessation service, barrel chest deformity, practical counselling, liquid carbohydrate-rich supplement, CAT test, blood gas, pseudomonas aeruginosa, acute change, prospective trial, exertional dyspnoea, poor adherence, copd postrehabilitation, sleep-disordered breathing, rapid tapering, single lung transplant, inspiratory airflow, long-acting bronchodilator, portable device, weight gain, bone density, cathepsin b, rehabilitation process, respiratory muscle fatigue, nicotine release, preoperative assessment, physiological fev1, ffm depletion, different inflammatory cell, arterial oxygen content, oxygen delivery devices and transtracheal, timed walk distance, breathing strategy, procedure from the diaphragm, haemophilus parainfluenza, endurance training, current cigarette smoker, volatile anesthetic, secretory leukoproteinase inhibitor, medical research council scale, onspecial unit, panlobular emphysema, normal lung, postoperative complication, life support, spontaneous skin bruising, systematic review, life supportive care, secretion clearance, perioperative management, physiological hypoventilation, selection criterion, tuberculosis onsetatallage, breath condensate, exacerbation-free interval, bilateral lung transplantation, sleep pattern, lung function, patient hospitalisation, procedure from the diaphragmthe, historical control, sleep-related disturbance, oxygen usage, clinical entity, atherosclerotic vascular disease, progressive dyspnoea, biomass fuel, o2-haemoglobin dissociation curve, severe hereditary deficiency, troponin, cardiac arrest, upright position, substantial benefit, sustained bronchodilitation, healthcare purchaser, surrogate decision-maker, CAT score, most mouth breather, regular ipratropium, nppv failure, peripheral muscle dysfunction, spirometric criterion, sterile sputum, radiographic feature, ventilatory assistance, cardiac surgery, dead space, slight hill, lung growth, inhalation technique, emphysema, coronary artery disease, average actuarial survival, family distress, abnormal growth, functional dyspnoea, cyanosisor bluish colour, tiotropium, home oxygen delivery, short-term oxygen therapy, alveolar ventilation/perfusion mismatching, arterial blood gas, smooth muscle, leukocytosis, o2 delivery setting, registered nurse, maximal lung function, impaired mental status, ventilatory muscle training, continuous flow dual-prong nasal cannula, inflammatory pattern, expiratory volume inone, short pulse, objective improvement, extreme fatigue, diagnostic problem, clinicians and healthcare delivery system, CT scanning, exercise flow setting, coliform, oral formulation, chest pain, chronic nppv, the icu, impaired health,
broad gold initiative, gas oxygen, high trapped lung volume, peripheral oedema, pathogenic mechanisms of copd, replacement therapy, a1-antitrypsin deficiency, loss off at mass, long-term oxygen therapy, intraoperative blood loss, local bacteria resistance pattern, nasal congestion, delivery of palliative care, intensive intervention, sleep oxygen flow rate, supplemental o2on request, respiratory drive, intensive care unit, broad category, life-threatening exacerbation, comfort adequate, chest radiography, advanced lung disease, spiritual domain, unit improvement in the tdi, fatal exacerbation, weight-stable copd patient, neurological disease, ecg change, airtrapping, respiratory tract, aspiration risk, maximal oxygen, life-threatening hypoxaemia, special respiratory care unit, collaborative self-management, lung resection, never smoked, therapeutic application, many exacerbation, ratio rv/tlc, copd, normal sa, oropharyngeal candidiasis, small transient fall, principle of palliative care, upper lobe predominance, mutual support, energy conservation, seizure risk, heart failure, excessive daytime sleepiness, end-of-life issue, copd diagnosis, p-value lvrs medical, endotracheal tube, pulmonary artery pressure, perioperative pulmonary complication, return of spirometric function, pulse oximetry sp, pulmonary vascular change, positive end-expiratory pressure, systemofpatient self-management, hospice service, respiratory stimulant, the pathogenesis, HRCT, intravenous infusion, past burden, prolonged intubation, central respiratory control, inflammatory cell profile, post rehabilitation exercise prescription, hyperacpnic, MRC dyspnoea scale, open thoracotomy, ventilator support, inspiratory rate, non-invasive ventilation, home support, comprehensive pulmonary rehabilitation, risk/benefit ratio, psychiatric condition, bioelectrical impedance analysis, emergency surgery, advance care planning, local oxygen, dyspnoea scale, effective palliative care, absolute contraindication severe musculoskeletal disease, human immunodeficiency virus, life expectancy, drug delivery, stationary delivery system, differential improvement, pronounced nocturnal oxygen desaturation, nonpurulent sputum improve, better-ventilated lung, sinus/ear pain, chronic health condition, nasal continuous flow, special care unit, infectious exacerbation, patient needtobereassessedwithin4week, pathological change, bronchodilator treatment, tension time index, spinal anaesthesia decrease, peak exercise capacity, unimportant finding, exacerbation history, vital capacity, proper precaution, exacerbator, vital sign, bacterial cellulitis, systemic inflammation, preventable threat, white cell count, acute hypercarbia, inspired o2, temporary decrement, case manager facilitator, patient mobility,
abundant proliferation, advance directive, breathing from a reservoir, litre flow, hormonal change, perioperative risk, pragmatic tip, mortality side effect, severe obesity, referral indication, small chamber, adequate pa, airway secretion, spirometric volume, bullous disease, suboptimal incorporation, lung infiltrate, serious freeze burn, ventilator-dependent patient, min 6-min walk distance, nonrebreathing face mask, chronic cough, vascular surgeon, heavy cigarette smoking, pulmonary dysfunction, inhalation exposure, family history of chronic respiratory illness, ciliary dysfunction, component drug, cardiac disease, mild exacerbation, patient satisfaction, death experience, ethical dilemma, bubble humidifier, white male smoker, patient self-management, nicotine nasal spray consist, pneumonia, alveolar attachment, enlarged gland, maximal exercise tolerance, medical therapy, tto patient, vascular smooth muscle, pathophysiological rationale, accurate prescription, prevalence of nutritional abnormality, maximal voluntary ventilation, selection algorithmare, gene expression, peak expiratory flow, transcription factor, cd8+ t lymphocyte, mucus ball, reactive specie, water heater, impaired growth, lung health study, muscle protein breakdown, low mortality rate, previous thoracic procedure, ventricular hypertrophy, impermeable backing, chest tightness, mdi administration, central airway, commuter plane, project benefit, electrocardiogram, gas exchange, nitric oxide, gradual progressive breathlessness, arterial oxygen tension, flow rate, rehabilitative element, diaphragm and tympanic sound, rheumatoid arthritis/fume exposure, inadequate blood handling technique, o2pressure threshold, deep breathing, shuttle walk distance, daytime pulmonary artery pressure, nutritional screening, comorbid illness, alpha1 antitrypsin, protein breakdown, very severe copd, dead space volume, extrathoracic organ, pulmonary rehabilitation process, cessation rate, response to inhaled corticosteroids, nocturnal desaturation, battery life, pharmaceutical agent, quality assurance, 12-lead ecg monitoring, LAMA, neutrophils, right ventricular failure, continuous administration, diaphragmatic breathing, rehabilitation programme, treadmill endurance, nasal cannula, additional o2supplementation, severe emphysema, airway obstruction from18, severe hyperinflation, definitive conclusion, chronic airways obstruction, childhood illness, direct spray, mucus plug, mild copd, oxygen flow, blood test, bronchial asthma, small reduction, ventilation/perfusion relationship, airway wall change, severe dyspnea, LABA, hiv infection, pathophysiological process, substantial percentage, antiprotease imbalance, specific requirement, cellular
oxygenation, cigarette smoke exposure, negative sputum, life sustaining intervention, maximum power output, intermittent positive-pressure breathing, therapeutic goal, effective primary prevention, daily living, daytime sleepiness, sleep deprivation, variable natural history, low birth weight, one-way valve, postoperative epidural analgesia, poor coping skill, normal lung function, physiological finding, macroscopic lesion, exertional capacity, localised bulla with vascular crowding, expiratory pressure, presence of comorbidity, tidal breathing, tobacco smoke, sleep apnoea syndrome, quality of well-being score, hydroxyl radical, nasal bridge ulceration, alveolar/arterial gradient, alveolar wall, marginal benefit, anatomic reservoir, aspiration pneumonia, O2 partial pressure, behavioural support session, bioelectrical impedance, post-surgical improvement, inflammatory cascade, informed decision, poor prognosis in exacerbation, breathing pattern, rate of lung function decline, acute episode, absolute none, utility of advance directive, low diffusion, accessory muscle contribution, light physical exertion, sleep problem, nicotine inhaler, drug interaction, breathing easy in anyway, serial ABG assessment, respiratory muscle function, low birth, a-trypsin augmentation therapy, O2 during exercise, COPD risk factor, hospital stay, expiratory airflow limitation, catheter insertion site, cessation treatment, around-the-clock coverage, persistent weaning failure, variable duration, disadvantages bulky, co-morbid disease, acute exacerbation, indoor air pollution, nocturnal polysomnography, normal circadian change, loss of cilia, peak oxygen uptake, reservoir cannula, severe respiratory acidosis, ambient temperature, cycle ergometry watt, tissue hypoxia, weight loss, heart rate, amelioration of patient, angiography vascular, lung volume reduction, one-year mortality, cosmetic obstructive adequate best, alveolar ventilation, secondary care, smoking cessation programme, nasal oxygen, nonrandomised trial, young adulthood, ability of the patient to cope with the environment, emotional function, FEV1/FVC, an inflammatory response, cough syncope, poor instrument calibration, undiagnosed initial phase, venturi mask, positive pressure ventilation, FEV1, environmental factor, ABG measurement, prolonged mechanical ventilation, kind permission, pulmonary vasoconstriction, daytime flow rate, non-nicotine drug, functional capacity, moderate-to-severe acidosis, end-of-life care, chronic disease, normal airway, inspired minute ventilation, office spirometry, continuous flow oxygen, small nasal inspiratory flow, severe coagulopathy, respective medical arm, NICE clinical guideline 12, ethnic group, cough rib fracture, conventional
mechanical ventilation, nonasal/ear irritation, noninvasive pulse oximetry, partial nicotine replacement, unstable patient, acidbase status, chronic respiratory impairment, low acuity of illness, congestive heart failure, gradual weaning, the hypoxia inhalation test, national emphysema treatment trial, cardiovascular disease, 4-min walk distance, respiratory impairment, short-acting agent, pulmonary embolus, prophylactic strategy, alveolar wall attachment, transtracheal team, tobacco, parenchymal destruction, cigarette smoker, pre-flight assessment, NICE clinical guideline 101, long acting bronchodilators, ICS, physician education, dermal surface, paradoxical movement, squamous cell carcinoma, male nonsmoker, quit date, exercise capacity and quality, severe airflow obstruction, troublesome symptom, presence of airflow limitation, bronchodilator reversibility, oedema, single lung transplantation, stand-by decision, adaptive decreased energy, defence mechanisms, bronchiectasis, mechanical ventilation, progressive disease, little bronchoreversibility, reputable healthcare, viral, body mass index, cigarette smoke, ERS 2004, oxidative stress, internal diameter, video-assisted thoracoscopic surgery, maximum exercise, neurological comorbidity, check inhalation technique, postoperative ventilator, mucous production, equivalent follow-up, leukotriene receptor antagonists/cromone, neutrophilia, lipid peroxidation product, life-supportive intervention, pulmonary vascular disease, ambulatory capability, elevated arterial carbon dioxide, patient tolerate oral medication, diabetes, bi-level positive airway pressure, dyspnea, low overall dyspnoea, ch wall deformity, mortality rate, surgery and lung transplantation, neck vein distension, diffuse panbronchiolitis effect, severity postbronchodilator, several clinical element, first-line pharmacological treatment, tobacco dependence exist, nose clip, homecare arrangement, clinically-meaningful threshold, low abdominal/pelvic surgery, protective inflammatory response, family history of respiratory illness, patient autonomy, vocal cord, stabilisation phase, nonhigh risk patient, slow wave, goblet cell metaplasia, pathophysiological change, procedure-related issue, connective tissue disease, elastic recoil, deep venous thrombosis, clinical trail, viscous secretion, lung surgery, dyspnoea, short hypoxic exposure, measure carboxyhaemoglobin, adoption rate, postbronchodilator fev1 support, transtracheal catheter, pulmonary function measurement, carbon dioxide retention, predominant ethical principle, oxygen administration, pulmonary hypertension, gastrointestinal problem, spirometric classification, peribronchial connective tissue, cardiac deconditioning,
psychosocial/behavioural intervention, airway inflammation, arterial blood gas analyser, alveolar/arterial o2 gradient, oxygen cost, copd mid-life onset, arterial sampling, preoperative arterial hypoxaemia, homogeneous emphysema and fev1, risk of toxicity, regional anaesthesia, airflow obstruction, emphysematous destruction, cardiopulmonary disease, spiritual issue, haemofillus influenza, respiratory muscle dysfunction, care plan, protein imbalance, small oxygen pulse, lung function decline, static lung volume, pulmonary rehabilitation, rapid endotracheal intubation, considerable variability, physiological abnormality, breathlessness, standard medical care, cardiopulmonary fitness, symptoms of sleep disturbance, progressive loss, maximum expiratory pressure, blood gas criterion, surgical intervention, socioeconomic status, routine assessment, gas exchange abnormality, chronic bronchitis, laparoscopic procedure, obstructive lung disease, environmental pollution, high-risk situation, heart disease, hoarse voice, nonconfrontational interaction, thoracic surgical procedure, profound deterioration, exercise capacity, cough, root decade, muscle strength, chronic relapsing disorder, respiratory infection, tracheobronchial cell, physiological indication, tidal volume, high-dependency respiratory unit, modern paradigm, authoritative source, irregular lung, bronchodilators, respiratory frequency, mucosal membranes may, extra- treatment social support, low mortality, severe disease gain solace, nicotine-containing cartridge, the 90-day surgical mortality, symptomatic merit, atmospheric air, pre-operative testing, adventitious rhonchus, unintentional weight loss, air travel hypoxaemia, effective analgesia, maintenance treatment, anticholinergic drug, arterial co-oximetry, symptoms of cough, system approach, harm reduction, moderate dyspnoea, abnormal enlargement, considerable benefit from pulmonary rehabilitation, normal dl, lung parenchyma, exercise-related energy expenditure, passive smoke exposure, avoid alcohol, maximal oxygen consumption, low effective dose, smoking behaviour, structural reorganisation, sleep quality, bupropion sr, intrinsic positive end expiratory pressure, body mass, pulmonary artery vasoconstriction, old age, nicotine patch, venous circulation, nasopharyngeal abnormality, supplementary material, readmission rate, muscle weakness, inspiratory capacity, anticipate challenge, oxygen cylinder, tumour necrosis factor-a, total lung capacity, invasive ventilation, addictive substance, H Flu, elevated pa, research laboratory, nocturnal pulmonary hypertension, oropharyngeal deposition, risk of pulmonary neoplasm, nebulised bronchodilator, chronic asthma, rescue
medication, physical examination, intermittent quitter, optimal disease management, young age, rapid inspiratory rate, long length, pursed-lips breathing, continuous smoker, concurrent cardiac disease, spirometry, acute care setting, peripheral neuropathy
Appendix B

Finer-grained COPD phenotypes

B.1 Problem

a) Condition

Chronic obstructive pulmonary disease (COPD)
Dyspnea
Cachexia/ muscle wasting/ loss of muscle mass
Exacerbations/Exacerbation frequency/Exacerbation severity
Non-exacerbator/ infrequent exacerbator
Exacerbator/ frequent exacerbator

Exacerbation phenotypes: bacterial, viral, eosinophilic, pauci inflammatory
(Chronic) bronchitis
Emphysema
broncholitis
asthma
Obesity
cardiovascular disease
diabetes, pulmonary hypertension
pulmonary vascular disease
congestive heart failure
myocardial infarction
stroke
Gastroesophageal reflux disease

b) RiskFactor
alpha1 antitrypsin
c-reactive protein (CRP)

i) SignOrSymptom
Shortness of breath/ Gradual progressive breathlessness
Expiratory airflow limitation/airflow limitation
Increased resistance of the small airways
Increased compliance of the lung
Parenchymal destruction
(Chronic) airways obstruction
Airways disease/ airway inflammation
(Chronic) cough
Sputum production/ purulent sputum
Pursed lips breathing
Hyperinflation
Sustained lung function/ rate of lung function decline (decrease)
Neutrophilic inflammation/ eosinophilic inflammation

ii) IndividualBehaviour
Tobacco exposure
smoking history
smoking status
pack years
smoking cessation
iii) Test Result
   - Increased white blood cell counts
   - Pulmonary artery/aortic diameter ratio

**B.2 Treatment**

- Supplemental oxygen/ oxygen therapy
- Pulmonary rehabilitation
- Bronchodilator response
- Enhanced response to inhaled corticosteroids

**B.3 Test**

a) Radiological Test
   - Computed tomography (CT) scanning/ HRTC

b) Microbiological Test
   - CBC

c) Physiological Test
   - Spirometry
   - FEV1/ Forced Vital capacity (FVC)/ FEV1/FVC ratio
   - Global Initiative for Chronic Obstructive Lung Disease (GOLD)
   - NICE clinical guideline 12/ NICE clinical guideline 101/ ATS/ERS 2004
   - MRC dyspnoea scale
   - COPD assessment test (CAT test/ score)
   - Hospital anxiety and depression questionnaire (HADS)
   - BODE index (BMI, Airflow obstruction, dyspnea, exercise capacity)