HYPHEN

DOI:

[10.1075/term.24.1](https://doi.org/10.1075/term.24.1)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Thompson, P., & Ananiadou, S. (2018). HYPHEN: A flexible, hybrid method to map phenotype concept mentions to terminological resources. *Terminology*, 24(1), 91-121. <https://doi.org/10.1075/term.24.1>

Published in:

Terminology

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

HYPHEN

A flexible, hybrid method to map phenotype concept mentions to terminological resources

Paul Thompson & Sophia Ananiadou

University of Manchester, UK

Narrative clinical records and biomedical articles constitute rich sources of information about *phenotypes*, i.e., markers distinguishing individuals with specific medical conditions from the general population. Phenotypes help clinicians to provide personalised treatments. However, locating information about them within huge document repositories is difficult, since each phenotypic concept can be mentioned in many ways. Normalisation methods automatically *map* divergent phrases to unique concepts in domain-specific terminologies, to allow location and linking of all mentions of a concept of interest. We have developed a hybrid normalisation method (HYPHEN) to handle concept mentions with wide ranging characteristics, across different text types. HYPHEN integrates various normalisation techniques that handle *surface-level* variations (e.g., differences in word order, word forms or acronyms/abbreviations) and *lexical-level* variations (where terms have similar *meanings*, but potentially unrelated *forms*). HYPHEN achieves robust performance for both biomedical academic text and narrative clinical records, and has the ability to significantly outperform related methods.

Keywords: normalisation, term variation, phenotypic information, terminological resources

1. Introduction

Phenotypic information consists of observable traits of living beings, which result from genetic make-up and/or environmental influences. Gathering information about different types of markers (such as signs and symptoms, risk factors and individual behaviours) that distinguish individuals with specific medical conditions from the general population is important in a range of scenarios, e.g., to

allow medical professionals to select appropriate therapies for individual patients (Han et al. 2010), to help patients better understand their diseases and treatments (Suominen et al. 2013), etc. In such scenarios, it is vital to facilitate access to as much relevant information as possible from complementary sources, e.g., individual patient details in narrative clinical records and summaries of research findings in biomedical articles (Patrick et al. 2007, Ruch et al. 2008). Although linking information in clinical records with other data sources is an active research area (Névéol and Zweigenbaum 2016), this can be extremely challenging, due to the many potential ways in which each concept of interest can be mentioned in text (Duclos et al. 2014). These include various types of linguistic and morphological variations (Rais and Grabar 2015), which change according to text type. For example, biomedical articles tend to include more technical language, often using terms derived from Greek or Latin to describe medical concepts (e.g., *hypercholesterolemia*, *dyspnea*, *leukocytosis*). In contrast, a more informal style is often adopted in clinical records, where brief narrative phrases or acronyms/abbreviations may describe concepts (e.g., *cholesterol elevation*, *could not breathe*, *high WBC*). This variability makes it virtually impossible to formulate search queries accounting for *all* possible ways in which a concept of interest could be mentioned in text, especially when searching across heterogeneous document types. This can inevitably lead to information overlook.

Ideally, search systems should allow *concept-level* search, rather than requiring users to try to enumerate the range of words and phrases that could potentially describe a concept in text. Existing domain-specific terminological resources can act as the starting point for this functionality. Such resources may be specialised for particular types of concepts, such as genes (Maglott et al. 2011), diseases (Schröml et al. 2012), clinical care (Donnelly 2006), phenotypes (Hamosh et al. 2005, Groza et al. 2015), or a combination of these (Bodenreider 2004), and typically consist of entries for a large number of concepts. Each entry includes various types of information, including textual variants or synonyms. These resources provide scope to automatically process documents to identify which words and phrases denote which domain-specific concepts, by matching them against the synonyms listed in the resources.

In practice, however, the situation is more complex. Firstly, terminological resources are typically manually curated by domain experts. Although this ensures high-quality content, their large size makes it virtually impossible to ensure that comprehensive information is consistently created and updated for every single concept. In the UMLS Metathesaurus (Bodenreider 2004), for instance, which covers a broad range of biomedical and health-related concepts, entries for some concepts include a wide range of possible textual variants. Variants for the phenotype concept *dyspnea* include spelling variations (*dyspnoea*),

terms with completely different forms (e.g., *respiratory difficulty*, *breathing difficulties*, *shortness of breath*), abbreviations (e.g., *SOB*) and terms with alternative internal structure (e.g., *breath shortness*). In contrast, the entry for *asthenia* lists no similar types of variations, making it difficult to establish that words and phrases such as *fatigue* or *little strength* actually refer to the same concept. Even when several variants *are* listed in the terminological resource, the creativity and unpredictability of language means that numerous subtle differences may occur amongst actual textual mentions, e.g., *respiratory difficulty* vs. *respiratory difficult ties* or *breathing difficulties* vs. *difficulties in breathing*. Although seemingly insignificant, unexpected variations of concepts may be problematic for automated methods.

To address such issues, various *normalisation* methods have been developed, which aim to facilitate mapping of textual concept mentions to appropriate concepts in terminological resources, even when no exact match with variants listed in the resource can be found. The majority of existing normalisation methods has focussed primarily on handling *surface-level* variations between textual mentions and listed concept synonyms, which use the same or related words (e.g., inflected, derived or abbreviated forms), possibly in different orders and with different syntactic structures. Whilst such techniques can handle a subset of the potential multitude of differences between concept mentions, it is important not to overlook *lexical-level* variations, where the similarity between mentions lies at the level of *meaning*, with possibly little, if any, surface-level similarity between phrases denoting the same concept. Handling such variations is particularly important when trying to link mentions of the same concept that occur in both biomedical articles and clinical records.

In this article, we describe a new hybrid method for normalisation of phenotype concepts (HYPHEN). It extends upon previous work on automatically recognising phenotype information in text (Fu and Ananiadou 2014, Fu et al. 2014, Fu et al. 2015), by allowing mentions of specific phenotype concepts found in both biomedical articles and narrative clinical records to be located and linked together. HYPHEN is designed to handle the wide range of surface and lexical level variations that occur amongst phenotype concept mentions in these text types. The flexible nature of HYPHEN means that it could be customised for application to other text types, and used to normalise biomedical concepts of different types to alternative terminological resources, with minimal effort.

HYPHEN integrates six separate normalisation techniques, each designed to handle a specific subset of variation patterns that occur amongst phenotype concept mentions. Techniques to handle lexical-level variations make use of two domain-specific knowledge resources to enable linking not only between mentions which vary according to the substitution of one of more semantically similar

words (*elevated* cholesterol vs. *raised* cholesterol), but also between formal, neo-classical medical terms and their more descriptive English equivalents (e.g., *kidney disease* vs. *nephrosis*). We combine these with techniques to handle surface-level variations, which facilitate mappings between plural and singular forms (e.g., *high blood sugars* vs. *high blood sugar*), terms with differing internal structures (e.g., *lung volume reduction* vs. *reduced lung volume*), and abbreviations/acronyms and their full forms (e.g., *RBS* vs. *respiratory distress syndrome*).

By demonstrating incremental increases in normalisation performance as the individual techniques are integrated in an optimal way to create a hybrid method, we illustrate the complementarity and necessity of combining individual techniques in constructing a robust, flexible and wide-coverage normalisation approach. We show that our hybrid system can achieve levels of performance that are either competitive with, or superior to, those achieved by related methods when applied to the tasks of normalising phenotype mentions in both biomedical abstracts and narrative clinical records.

2. Related work

A wide range of methods has been developed to handle *surface-level* variations between textual concept mentions and synonyms listed in terminological resources. Methods include removing inflections (e.g., plurals), ignoring certain non-matching words, generating derived words (e.g., *elevate* → *elevation*), or generating permutations of words in concept synonyms (e.g., *increase in blood pressure* → *blood pressure increase*), e.g., (Dai et al. 2008, Hersh and Greenes 1990, Jonquet et al. 2009, Savova et al. 2010, Zhou et al. 2006).

Several tools use terminological resources in conjunction with one or more surface-level normalisation techniques to jointly perform concept recognition *and* normalisation (Aronson and Lang 2010, Nunes et al. 2013, Tanenblatt et al. 2010, Jonquet et al. 2009). Since these tools are driven by information contained within the resource, only those phrases that resemble variants listed in the resources will be recognised as concept mentions; incomplete coverage of variants means that many concept mentions may fail to be recognised and resolved to appropriate concepts.

An alternative approach to concept recognition uses machine learning (ML) to teach tools how to recognise phrases denoting concepts of different types, e.g., (Bodnari et al. 2013). Typically, ML techniques learn using annotated corpora, i.e., document collections in which human experts have manually marked up concepts. Annotated corpora relevant to phenotype recognition include those consisting of biomedical abstracts, e.g., (Leaman et al. 2009, Doğan et al. 2014) and

clinical records, e.g. (Uzuner et al. 2011, Pradhan et al. 2015). ML methods usually learn high-level linguistic patterns; they are often independent of the exact words used to describe concepts, and may have little, if any, reliance on information in terminological resources to determine which phrases constitute concept mentions.

An advantage of ML-trained tools is that they can recognise a far wider variety of concept mentions than tools constrained by the contents of terminological resources. However, most annotated corpora do not include expert-added mappings to concepts in terminological resources. Accordingly, most ML-trained concept recognition tools are designed only to *recognise*, but not to normalise, concepts of interest. Whilst existing normalisation methods can be applied to the output of ML-recognised concept mentions, the high degrees of lexical and structural variability amongst the automatically recognised terms can present significant challenges (Oellrich et al. 2015).

To encourage the development of more robust methods of recognising *and* normalising widely divergent concept mentions, several corpora of either biomedical abstracts or clinical records have been released with expert-added links between annotated concept mentions and appropriate concept entries in terminological resources, e.g., (Doğan et al. 2014, Leaman et al. 2009, Suominen et al. 2013). These corpora provide important evidence about possible variations amongst concept mentions in different text types. They have stimulated novel approaches to normalisation, particularly in the context of shared tasks, in which different research groups develop solutions to common problems (Pradhan et al. 2014, Elhadad et al. 2015, Pradhan et al. 2015).

However, given that each of the above corpora only includes documents belonging to a single type, many normalisation methods are specialised for application to *either* biomedical text or clinical records, but not both. An additional issue is that the different corpora define mappings to different terminological resources, which include MeSH, SNOMED-CT (Donnelly 2006), and the UMLS Metathesaurus (Bodenreider 2004). Until recently, therefore, it has been problematic to develop methods that can robustly normalise concept mentions in both text types, and link these mentions by mapping them to a common terminological resource. However, the recent emergence of corpora of documents drawn from multiple text types, with mappings to a common terminological resource, e.g., (Alnazzawi et al. 2016, Wang et al. 2016) provides new opportunities to develop and evaluate wider-coverage, flexible normalisation tools.

Promising recent normalisation approaches have combined/ranked results from several different methods (Collier et al. 2015), or used pattern-matching to handle frequently occurring variations not listed in the terminological resource (e.g., (Fan et al. 2013, Ramanan et al. 2013, Wang and Akella 2013). The availability

of annotated corpora that include concept mappings has paved the way for ML-based normalisation approaches, which have demonstrated high performance when applied both to biomedical abstracts (Leaman et al. 2013) and to clinical records (Leaman et al. 2015). Whilst ML methods alleviate the need to manually develop techniques or patterns to map between different types of concept variants, a potential disadvantage is their typically high sensitivity to the features of the text on which they were trained, which can limit their utility for wide-scale application across multiple text types/domains.

String similarity methods (e.g., (Kate 2015, Dogan and Lu 2012) offer a flexible means of establishing matches between concept variations, which, similarly to ML-based methods, do not require extensive human input to define variation patterns. Word-level similarity metrics (Jaccard 1912) can ensure that a match between mentions is only considered if a certain proportion of words is shared between them. Character-based similarity methods (Cohen et al. 2003) can relax the constraints, such that words that closely resemble each other are considered to be equivalent. They can link concept mentions that use different inflected or derived forms of words, or which contain spelling errors (a common feature of narrative clinical records (Patrick et al. 2007)).

In (Alnazzawi et al. 2016), word and character-based string similarity metrics are integrated in the PhenoNorm method, allowing matching amongst phenotype concept mentions with a wide variety of internal variations. These include differing numbers and/or orders of words (*increasing chest pain* vs. *chest pain increasing in severity*) and different derivations of words (*hyperkalemia* vs. *hyperkalemic*). However, in some cases, this technique is *too* flexible, e.g., PhenoNorm incorrectly maps *chronic leg edema* to *chronic leg ulcer* instead of *leg edema*, according to greater string-based similarity, despite the differences in meaning. PhenoNorm also integrates functionality to handle lexical-level variations, by using a general language lexical resource, i.e., WordNet (Fellbaum 1998), to find synonyms of each noun and adjective occurring within concept mentions. All possible lexical-level variations of the original concept mention are then systematically generated. This allows, e.g., *worsening renal function* to be mapped to *decreased renal function*. However, since such variations are generated on a word-by-word basis, PhenoNorm is unable to map the mention *increased oxygen requirement* to the concept *hypoxia*. Despite these issues, however, PhenoNorm achieves robust performance in normalising different concept types in different text types, without training or reconfiguration.

3. Methods

The HYPHEN method attempts to address some of the issues and challenges of normalisation highlighted by previous efforts. To handle as many potential variations amongst phenotype concept mentions as possible, we employ six different individual normalisation techniques, each designed to handle a different class of either surface or lexical-level variations. We apply these techniques either by implementing/extending upon ideas that have been introduced elsewhere, or by using existing tools. All techniques *generate* potential variants of phenotype concept mentions. If the original concept mention cannot be matched against a synonym in the chosen terminological resource, but one of the generated variants can achieve a match, then a mapping is established. Each technique generates variants in a systematic way, aiming to ensure that all such variants are semantically consistent with the original mention, to keep incorrect mappings to a minimum.

In this section, we firstly describe each of the different individual normalisation techniques, and then explain how we combine them into a hybrid method in an optimal way, to best take advantage of the complementary nature of the output of each technique.

3.1 Lexical variant generation

We employ three approaches to generating lexical-level term variants. The first is aimed at generating descriptive English equivalents of neoclassical compound medical terms, whilst the second works in the opposite direction. Given that these two different types of terms usually bear no surface level resemblance to each other, and since one or other of the term types may be missing from terminological resources, this technique is important to allow links to be established effectively between information in the literature and clinical records for individual patients.

The third method is largely inspired by the word-by-word lexical-level synonym generation technique used in PhenoNorm. However, we use the UMLS Metathesaurus as our lexical resource, rather than WordNet, with the aim of generating a greater number of domain-relevant alternatives.

3.1.1 *Transformation between English and Neoclassical terminology*

Hippocratic writings in Greek, dating from the 5th and 4th centuries BC, are the oldest written sources of western medicine. They are considered to constitute the beginning of the Greek era of the language of medicine (Wulff 2004), which has been used ever since. Whilst Latin was subsequently introduced as a further source of formal medical terms, many novel concepts introduced by medical sci-

entists have been traditionally composed of Greek rather than Latin roots, partly because Latin does not permit the formation of composite words to the same extent (Wulff 2004). For example, the Greek terms *nephrectomy*, *ophthalmoscopy* and *erythrocyte* are less cumbersome terms than their Latin equivalents, i.e., *excisio renis*, *inspectio oculorum* and *cellula rubra*, respectively.

The huge stock of Greek medical terminology also presents 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)). Many Greek prefixes and suffixes are also more productive than Latin ones. For example, the Greek prefix “hyper-” is more productive than Latin “super-”, despite their comparable meanings. As a result, *hypertension* is used, rather than *supertension*. However, both Greek and Latin terms are still widely used in the medical domain; compound terms composed of forms that originate from these two languages are known as *neoclassical compounds*. Nowadays, however, as English has become the language of choice of medical scientists for international communication, neo-classical terminology is often used interchangeably with phrases of equivalent meaning, but composed of common English words. These alternative phrases are particularly frequent in narrative clinical records, while more formal terminology is still prevalent in academic articles. Examples of these very different ways of referring to the same concepts are shown in Table 1.

Table 1. Examples of phenotype concepts expressed as neoclassical and English terms

Neoclassical concept mention	Equivalent English concept mention
<i>Aphasia</i>	<i>language dysfunction</i>
<i>Apnea</i>	<i>inability to breath</i>
<i>Dysphagia</i>	<i>difficulty swallowing</i>
<i>Dyspnea</i>	<i>shortness of breath</i>
<i>hyperglycemia</i>	<i>elevated blood sugar</i>
<i>hypertension</i>	<i>high blood pressure</i>
<i>thyromegaly</i>	<i>enlarged thyroid</i>

A number of studies have defined linguistically grounded methods for morphological analysis of medical neoclassical compounds (Deléger et al. 2007, Namer and Baud 2005, Ananiadou 1994). In a more practical application, i.e., cross-language information retrieval in the medical domain (Markó et al. 2005), it was noted that decomposing complex medical terms into sub-words, and establishing translation equivalences between these sub-words rather than complete terms, could result in improved retrieval performance. Our approach follows sim-

ilar ideas to (Markó et al. 2005), since we also decompose compound medical terms into “sub-words”. As described in (Ananiadou 1994), there are certain rules governing the formation of neoclassical compounds – they consist of roots, which optionally or obligatorily have affixes attached to them to create compounds. Usually, roots and affixes have their own special meanings, and the position of specific affixes is fixed, i.e., before the root as a prefix, or after the root as a suffix. This information can help in the normalisation of concept mentions in two main ways. For mentions that constitute neoclassical compounds, identifying their component parts and determining the possible meanings of each of these elements can allow the generation of potential English multi-word phrases that have equivalent meanings to the compound terms. Conversely, for mentions that consist of English words, the ability to determine equivalences between English words and elements of neoclassical compounds *and* how these elements can be combined together to create valid neoclassical compounds, makes it possible to generate possible neoclassical equivalents of more informal concept mentions.

Information about the combinability of neoclassical elements helps to ensure that only valid compounds are generated, in order to increase the efficiency of the variant generation process. For example, English phrases that describe swallowing disorders may include *difficulty in swallowing* and *swallowing difficulties*, i.e., the ordering of the words may vary. We can map the words in these terms to elements of neoclassical compounds (i.e., *swallowing* to *phagia* and *difficulty* to *dys*). Using the information that *dys* can only appear in prefix position, we can determine that *dysphagia* is a valid neoclassical compound, but *phagiadys* is not.

According to the above, our method is driven by a specialised terminological resource. This consists of roots and affixes used in the formation of medical neoclassical compounds, along with information about valid positions of each element within these compounds, and accompanied by one or more potential English equivalents. This resource is based largely on the information provided in an online listing,¹ consisting of 809 neoclassical roots, prefixes and suffixes used in medical terms, together with equivalent English words or descriptive characterisations. The listing also provides information about valid positions of each element within compounds. We transformed certain aspects of the original inventory, to create a consistent format that can be accessed and processed by our methods. Some example entries in the resulting terminology are shown in Table 2 (the dashes in the neoclassical elements indicate whether they can appear at the start or end of a compound).

1. List of medical roots, suffixes and prefixes, Wikipedia (https://en.wikipedia.org/wiki/List_of_medical_roots,_suffixes_and_prefixes). Accessed 8 February 2018)

The procedure for generating neoclassical compounds from English descriptive phrases (E2N) is illustrated in Figure 1. For *difficulty breathing*, “dys-” is retrieved as a prefix with the meaning “difficult” while “-pnea, -pnoea” are found as alternative equivalents for “breathing”. Using the positional information in the resource, we can combine the elements in the correct order to create valid compound terms, i.e., *dyspnea* and *dyspnoea*.

Table 2. Examples of neoclassical elements and their English equivalents

Neoclassical element	Equivalent English expression
abdomin-/ abdomino-	abdomen, abdominal
cyt-/ cyto-/ -cyte	Cell
-dynia	Pain
-esophageal/ -esophago	gullet
gastr-/ gastro-	stomach
hepat-/ hepatic-	liver
-ismus	spasm, contraction
kin-/ kine-/ kino-/ kinesi-/ kinesio-	movement
leuc-/ leuco-/ leuk-/ leuko-	white
nerv-/ neur-/ neuro-/ neuro-	nerves, nervous system
olig-/ oligo-	little, few
xanth-/ xantho-	yellow, abnormally yellow

Figure 2 illustrates the inverse process, i.e., generating English descriptive equivalents of neoclassical compounds (N2E). The resource is firstly used to decompose *dyspnea* into two different elements (i.e., “dys-” and “-pnea”). The corresponding English translations are then retrieved (i.e., “difficult, abnormal, failed”, etc. for “dys-“and “breath, breathing” for “pnea”) and possible multi-word phrases are generated in a systematic manner (e.g., *difficult breath*, *difficult breathing*, *abnormal breath*, *abnormal breathing*, etc.).

Further examples of composition and decomposition between English and neoclassical terms are shown in Table 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, possible English equivalents of this term are *without strength* and *without power*. Furthermore, since *asthenia* as a whole could also be translated as *weakness* or *fatigue*, we also generate these terms as possible variants. For English terms with more than two words, such as *high blood sugar*, we generate multiple variants. Both *hyper-*

Figure 1. Example of the formation of neoclassical phenotype terms

Figure 2. Example of the formation of English phenotype terms

and *glyc-* are prefixes, so we generate both *glychyperemia* and *hyperglycemia*. Although the former term is not valid, this does not matter, as only the latter term *hyperglycemia* will achieve a match in the terminological resources.

Table 3. Examples of transformation between English and neoclassical concepts

(a) English to neoclassical		
Original English concept	Corresponding elements	Generated neoclassical concept
<i>failed swallow</i>	failed: dys-; swallow: -phagia.	<i>dysphagia</i>
<i>high blood sugar</i>	high: hyper-; blood: -emia; sugar: glyc-.	<i>hyperglycemia</i>
<i>without strength</i>	without: a-; strength: -sthenia.	<i>asthenia</i>
<i>thyroid enlargement</i>	thyroid: thyro-; enlargement: -megaly.	<i>thyromegaly</i>
<i>increased white blood cells</i>	white, white blood cells: leuko-; cell: cyt-; increased: -osis.	<i>leukocytosis</i>
(b) neoclassical to English		
Original Neoclassical concept	Elements and their meanings	Generated English concept(s)
<i>dysphagia</i>	dys-: failed, difficult; -phagia: eat(ing), swallow(ing).	<i>failed swallow(ing), failed eat(ing), difficult swallow(ing)...</i>
<i>hyperglycemia</i>	hyper-: (abnormally) high, elevated; glyc-: sweet, sugar; -emia: blood (condition).	<i>high sugar blood, high blood sugar, elevated blood sugar...</i>
<i>Asthenia</i>	a-: without, loss of; -sthenia: strength, power; -asthenia: weakness, fatigue.	<i>without strength, loss of strength, weakness, fatigue...</i>
<i>thyromegaly</i>	thyro-: thyroid; megaly: enlargement.	<i>thyroid enlargement</i>
<i>leukocytosis</i>	leuko-: white, white blood cells; cyt-: cell; -osis: increased, disease, condition.	<i>increased white blood cells</i>

3.1.2 Synonym searching

Our third lexical-level variant generation method aims to account for the high variability amongst English phrases that can describe medical concepts. For example, the condition *hypercholesterolemia* could also be described more simply using the phrases *elevated cholesterol* or *high cholesterol*. For more complex concepts like *moderate left ventricular systolic dysfunction*, a potential multitude of similar phrases could represent the same concept. For example, *ventricle* could replace *ventricular*, *systole* could replace *systolic* and *malfunction* is an alternative for *dysfunction*, amongst others. However, it is rare for all possible combinations to be listed in terminological resources. In the UMLS Metathesaurus, for example, *elevated cholesterol* is a synonym of *hypercholesterolemia*, but *high cholesterol* is

not. For *moderate left ventricular systolic dysfunction*, however, there are no synonyms that include alternative words.

This situation motivates the type of compositional (i.e., word-by-word) generation of synonyms that was employed in the PhenoNorm method. We follow the same general idea, although we employ the domain-specific UMLS Metathesaurus instead of the general language WordNet, in order to ensure more relevant synonyms. For example, *high* appears as a synonym of *elevated* in the Metathesaurus, but not in WordNet; the same is true of *malfunction* as a synonym of *dysfunction*.

3.2 Syntactic normalisation

As is described in (Jacquemin 1999), variations amongst descriptive mentions of concepts may involve not only the substitution of semantically similar words, but also differences in the internal structures of terms. These can include changes in the order of the words and/or the use words belonging to the same morphological family, but having different parts of speech. For example, the phrases *abdomen pain*, *pain in abdomen*, *abdominal pain* and *painful abdomen* all describe the same concept.

Similarly to (Jacquemin 1999), we take a systematic approach to the generating syntactic-level variants of concept mentions that have structures such as the above, i.e., we match terms with particular syntactic structures and use rules to transform them into terms with equivalent meaning, but with different structures. However, in contrast to (Jacquemin 1999), whose variant generation process is reliant on hand-crafted grammar rules, we use the probabilistic head-driven phrase structure grammar (HPSG) parser Enju (Miyao and Tsujii 2008) to obtain the predicate-argument structures of concepts, which capture the relationships between head and dependent words (see Table 4).

Table 4. Examples of different syntactic constructions of concepts identified by Enju

Category	Name	Concept	Predicate	Argument	
				Argument 1	2
C1	noun_arg1	<i>abdomen pain cholesterol elevation</i>	<i>Abdomen cholesterol</i>	<i>pain elevation</i>	
C2	prep_arg12	<i>pain in abdomen elevation of cholestetrol</i>	<i>in of</i>	<i>pain elevation</i>	<i>abdomen cholesterol</i>
C2	adj_arg1	<i>painful abdomen abdominal pain elevated cholesterol</i>	<i>painful abdominal elevated</i>	<i>abdomen pain cholesterol</i>	
C4	verb_arg1	<i>cholesterol elevated elevating cholesterol</i>	<i>elevated elevating</i>	<i>cholesterol</i>	

If a concept occurs in text in any of the forms C₁, C₂, C₃ or C₄, then variants of the other three forms are automatically generated using rules. These are illustrated in Figure 3. The different derived word forms necessary for these transformations (e.g. *elevation* -> *elevated*, *pain* -> *painful*) are obtained using a combination of information from the CatVar database (Habash and Dorr 2003), which contains categorical variants of English lexemes, and the domain-specific BioLexicon resource (Thompson et al. 2011). The inclusion of noun plurals in CatVar can also be useful in mapping mentions that appear in the singular form in concept mentions, but only in the plural form the terminological resource, e.g. *cardiac disease*, which should be mapped to the concept Doo6331 *cardiac diseases* in MeSH.

Figure 3. Syntactic transformation of *cholesterol elevation*: C₁ to C₂; C₁ to C₃; C₁ to C₄

3.3 Acronym and abbreviation disambiguation

Acronyms and abbreviations can be challenging targets for normalisation. Firstly, terminological resources may not consistently list acronyms/abbreviations within the synonym set for each concept. For example, *BGL* is not listed as a synonym of *blood glucose level* in the UMLS Metathesaurus. Furthermore, many acronyms and abbreviations can have several possible interpretations, which can be context-dependent (Carroll et al. 2012). For example, although *WBC* means *white blood cell* in many medical documents, it also has other possible expansions, including *Well Baby Clinic* and *warm blood cardioplegia*. To handle this, we make use of the previously developed Acromine Disambiguator tool² (Okazaki et al. 2010), which

2. *Acronime disambiguation demonstration* (http://www.nactem.ac.uk/software/acromine_disambiguation/). Accessed 8 February 2018.

automatically determines the most appropriate full form for acronyms and abbreviations occurring in text, based on their context. The tool has been shown to achieve high performance on MEDLINE abstracts (0.986 F-score). Since acronym disambiguation and expansion are based on information obtained by processing the whole of MEDLINE, Acromine Disambiguator can determine interpretations of a much wider range of acronyms than would be possible by relying only on a manually-curated terminological resource. Furthermore, even for those acronyms that *are* listed in the terminological resource, the context-sensitive interpretations produced by Acromine Disambiguator help to ensure that ambiguous acronyms are automatically mapped to the most appropriate concept.

3.4 Plural to singular

A final method that is used elsewhere is the replacement of concept mentions appearing in the plural form with their singular equivalents. This is important, since many concept synonyms in terminological resources are listed only in the singular format, whereas in documents, both singular and plural forms may be encountered. We use the Enju parser to generate singular variants, e.g., to convert *thrombi* into *thrombus*.

3.5 Hybrid methods

Hybrid systems, which combine the results of a number of methods, have been shown to achieve superior performance to the individual methods from which they are constructed, e.g., (Kang et al. 2010, Uzuner et al. 2011). We therefore hypothesised that a hybrid method should produce superior results to those achieved by any of our individual techniques, which are each specialised to handle only a subset of possible variations that can occur amongst concept mentions. Only by applying a range of these techniques can we ensure that as many concept mentions as possible are successfully normalised. Furthermore, we found that a concept mention must sometimes undergo various stages of transformation to allow successful mapping to a synonym in the underlying terminological resource. For example, an acronym may firstly have to be expanded to its full form, which itself may have to undergo further transformations before it can be matched against a concept synonym in the resource. Accordingly, as shown in Figure 4, our hybrid HYPHEN system is created by combining the individual methods into linear chains.

In HYPHEN, the individual normalisation techniques are ordered in an optimal way for a given dataset. We have performed experiments to determine the best ordering for each dataset. After each method has been run over the set of

Figure 4. Combining individual normalisation methods

concept mentions, it is determined whether any of the generated variants matches a concept synonym in the chosen terminological resource (e.g., MeSH); if so, then the hybrid method terminates. Otherwise, the output of the method is used as input to the next method in the sequence, and a match is sought amongst these. In some cases, the first method may cause the original concept mention to be transformed, while in others, it may remain unchanged. This behaviour is designed to increase the chances of normalising concept mentions in scenarios such as the one described above, where original concept mentions can only be mapped to a concept in the terminological resource after undergoing multiple transformations. The process continues until a match is found, or until all techniques combined in the hybrid method have been applied.

4. Results

To ensure cross-domain applicability, we have evaluated our methods on gold-standard corpora consisting of different text types, i.e., clinical records and biomedical abstracts. Specifically, we used the 299 clinical records annotated with disorder mentions normalised to concept unique identifiers (CUIs) in SNOMED-CT, released in the context of the ShARe/CLEF eHealth Evaluation Lab 2013 (Suominen et al. 2013), and the 1,500 PubMed abstracts annotated with disease mentions normalised to MeSH IDs, created for the BioCreative V Chemical Disease Relation (CDR) Task (Li et al. 2016). The numbers of gold standard, normalised phenotype annotations in each corpus are 11,167 (3,821 unique) and 18,797 (3,737 unique), respectively.

Table 5 shows the number of variants generated by applying each of our individual methods to all annotated concept mentions in each of the corpora. There are considerable differences between the numbers of variants generated by some of the different methods, and sometimes between the different corpora.

Generally, the number of variants generated for clinical records is considerably larger than for abstracts, which can be explained by the greater use of more variable English descriptive phrases in clinical records. This, together with the large number of terms and synonyms listed in the UMLS Metathesaurus, helps to explain the huge number of variants generated by the SS method for clinical records. While in this text type, qualifier terms such as *high*, *low*, *elevated*, *reduced* and *decreased* may be used interchangeably in different mentions of the same concept, such ideas are more likely to be conveyed within compound technical medical terms using prefixes such as *hyper* or *hypo*, and hence the productivity of the SS method is lower for abstracts. SN also generates many variants, due to the various transformations performed on each term, combined with the different possible morphological variants provided by CatVar and the BioLexicon. A fairly large number of variants is found by N2E, since each neoclassical affix can typically be translated using several English words. The number of variants generated by this method for clinical records is almost as high as for the biomedical abstracts, which, perhaps unexpectedly, provides evidence that complex medical terms are widely used in both text types. Conversely, E2N generates far fewer variants, partly because several English words are often mapped to a single neoclassical compound element, and because the ordering of the elements within the resulting compound term is relatively fixed. The number of E2N variants is highest for clinical records, which reinforces our observation that English phrases are more prevalent in this text type than in abstracts. The plural to singular normalisation (P2S) and the acronym and abbreviation disambiguation (AAD) techniques generate the smallest number of variants – the only function of P2S is to generate singular forms from plurals, while AAD determines a single, disambiguated full form for each acronym or abbreviation in text.

Table 5. No. of unique variants generated using different normalisation methods

Normalisation Methods	No. of unique variants generated	
	Clinical Records	Biomedical Abstracts
Synonym searching (SS)	243,372	112,126
English to neoclassical (E2N)	2,963	1,572
Neoclassical to English (N2E)	4,295	4,554
Syntactic normalisation (SN)	19,272	10,293
Acronym and abbreviation disambiguation (AAD)	321	451
Plural to single (P2S)	1,292	844
Total	271,515	129,840

4.1 Evaluation metrics

We report the performance of our individual techniques and hybrid methods in terms of precision, recall and F-Score. These metrics require that true positive (TP), false positive (FP) and false negative (FN) cases are firstly determined. They are then calculated as follows:

$$\text{F-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{Where Precision} = \text{TP} / (\text{TP} + \text{FP}), \text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

In our context, we calculate TP, FP and FN as follows:

1. TP – the concept mention or one its generated variants is found in the appropriate terminological resource (i.e., SNOMED-CT or MeSH) AND the mention is assigned the correct concept identifier, according to the gold standard (i.e., either a SNOMED-CT CUIs or MeSH Unique ID).
2. FP – the concept mention or one of its generated variants is found in the appropriate terminological resource and is assigned a concept identifier, BUT this concept identifier is incorrect according to the gold standard.
3. FN – neither the concept mention nor its generated variants can be found in the appropriate terminological resource BUT an identifier should have been assigned, according to the gold standard.

4.2 Baseline and individual methods

We compare the output of our methods against two baselines: *dictionary lookup*, where we determine whether each original concept mention matches exactly with a term in the appropriate terminological resource (i.e., either SNOMED-CT or MESH); and the widely-used MetaMap tool (Aronson and Lang 2010). In Table 6, the results of the baseline methods are compared with those obtained through application of the six individual normalisation methods on both the clinical record corpus and the biomedical abstract corpus. The statistical significance of the difference between the *dictionary lookup* baseline and our six individual methods is also indicated. We used bootstrapping to carry out 1000 runs of random replacement of concept mentions, and performed unpaired t-tests to obtain a P-value. We show with an asterisk cases in which the difference is considered to be significant ($p < 0.01$). Methods that outperform both baselines are emboldened.

In terms of F-scores, all individual methods outperform the dictionary lookup baseline, demonstrating that each is able to generate useful variants for improving mapping accuracy. In almost all cases, the improvement over this baseline is statistically significant, the only exception being the N2E method when applied to the clinical corpus. This is likely to be due to the larger size of SNOMED-CT vocab-

Table 6. Normalisation performance of individual methods compared to baselines (P = Precision; R = Recall; F = F-Score)

Method	Clinical (11,167 mentions)				Biomedical (18,797 mentions)			
	P	R	F	P-value	P	R	F	P-value
Dictionary lookup	0.6642	0.6020	0.6316	–	0.9621	0.6475	0.7741	–
MetaMap	0.5394	0.8781	0.6682	–	0.9069	0.7251	0.8030	–
Synonym Searching (SS)	0.6678	0.7605	0.7118*	<0.0001	0.9172	0.7748	0.8401*	<0.0001
Neoclassical-English (N2E)	0.6536	0.6115	0.6319	0.1054	0.9516	0.6536	0.7749*	<0.0001
English-Neoclassical (E2N)	0.6639	0.6036	0.6323*	0.0002	0.9613	0.6483	0.7745*	0.0041
Syn Normalisation (SN)	0.6639	0.6223	0.6424*	<0.0001	0.9577	0.6553	0.7781*	<0.0001
Acr./ Abbr. disambig (AAD)	0.6775	0.6454	0.6610*	<0.0001	0.9587	0.7163	0.8198*	<0.0001
Plural-Singular (P2S)	0.6431	0.6539	0.6484*	<0.0001	0.9620	0.6569	0.7807*	<0.0001

ulary, which reduces the effect of the method, since this resource already contains many neoclassical terms. The F-scores achieved by MetaMap are higher than most individual methods, although this is to be expected, given that MetaMap itself integrates a range of different techniques. Even so, several of the individual techniques are able to outperform MetaMap. For both corpora, SS considerably outperforms MetaMap, which highlights the importance of considering word-level synonyms. For example, in the biomedical corpus, the concept mention *cardiac malformations* should be mapped to the MeSH concept Do18376, whose only listed synonyms are *cardiovascular abnormalities* and *abnormalities cardiovascular*. However, the SS method generates *malformations* as a synonym of *abnormalities*, allowing the correct mapping to be achieved. Likewise, in the clinical corpus, the mention *ethanol intoxication* cannot be mapped to the SNOMED-CT concept C0001969 without finding *alcohol* as a synonym of *ethanol*.

The AAD method also outperforms MetaMap for the biomedical corpus, and approaches MetaMap's performance for the clinical corpus. This is likely to be due to the high occurrence of acronyms in both corpora, together with our employment of the sophisticated AcroMine Disambiguator tool, which considers both local (document-level) and global abbreviations (which may not be expanded in the same document). In contrast, MetaMap employs a simpler approach, in which a bracketed acronym must be preceded by its expansion, in the same sentence.

Accordingly, only AcroMine Disambiguator can disambiguate *ad* to *Alzheimer's disease* and *elevated icp* to *elevated intracranial pressure*. The lower improvement of this method over the baseline for clinical text is likely to be due to the less formal nature of the text, resulting in less standardised abbreviations than those used in biomedical text, and additionally since AcroMine Disambiguator was trained on biomedical abstracts. Furthermore, challenging ad-hoc abbreviations have been noted as a feature of the ShARe/CLEF corpus (Alnazzawi et al. 2016).

Both E2N and N2E have a rather less significant contribution than the other methods. One issue with E2N is that the fine-grained nature of the concepts in SNOMED-CT and MeSH means that some terms considered as equivalent by our methods (e.g., *asthenia* and *fatigue*) actually correspond to different concept identifiers in the resources, according to the subtle differences in their interpretation. However, there is evidence that E2N and N2E *do* produce a reasonable number of valid variants, even if they cannot help in our current evaluation. For example, although the MeSH concept D002637 only lists a single variant, i.e., *chest pain*, our E2N method generates the variants *thoracalgia*, *thoracodynia* and *throactic pain*, which all have the same meaning. Similarly, N2E can generate *joint inflammation*, which is a valid variant of *arthritis*, even though this variant is not listed in MeSH.

All of our methods achieve higher precision than MetaMap, and, for biomedical abstracts, precision is mostly extremely high (generally over 0.95). This provides evidence that our systematic methods are working as intended, i.e., most generated variants are semantically consistent with the original mention. The noticeably lower precision of SS is likely to be due to the huge size of UMLS, with vocabulary drawn from various different sources, meaning that it can be difficult to maintain complete semantic consistency amongst all listed synonyms. Although, with the exception of SS, the recall levels of our techniques are mostly lower than MetaMap, our individual methods in isolation are not intended to resolve all normalisation issues, and we expect recall to improve when the techniques are combined into a hybrid method.

The results for the clinical records are rather different from abstracts – precision levels are considerably lower, and recall is also usually slightly lower. Such results are not unexpected, since the wide and unpredictable variability amongst concept mentions in this text type presents significant challenges for our systematic variant generation methods. However, all of the individual methods perform better than the dictionary lookup baseline, demonstrating that they are all effective in generating valid variants, even in this challenging context.

4.3 Hybrid methods

To create our hybrid methods, we firstly carried out experiments to apply pairs of methods to both the clinical and biomedical corpora, to validate that the process of combining methods is actually advantageous. A fixed aspect of the hybrid methods is that the AAD method must be applied prior to other methods, because AcroMine Disambiguator requires concept mentions to be processed in their original textual context. Accordingly, we evaluated the effect of applying each of the other five methods after AAD, on both corpora. The results are shown in upper half of Table 7, where P-values are calculated in comparison to the AAD method alone. We show with an asterisk all cases where the performance gain achieved by adding the additional method to AAD is statistically significant ($p < 0.01$). The complementary nature of the variants generated by different individual methods is reflected in the fact that all pairings achieve statistically significant performance improvements compared to AAD alone, with the exception of N2E on clinical records. In both corpora, the greatest improvement was achieved by applying the SS method after AAD, which underlines the importance of incorporating methods to handle surface *and* lexical-level variations in each text type. The integration of these two methods is particularly strong for biomedical abstracts, where the combined performance is considerably higher than when either method is used in isolation. The precise nature of the output of AAD also helps to offset some precision issues of SS when it is used in isolation, since acronym expansions listed in the UMLS Metathesaurus may not be as comprehensive or accurate as those that can be assigned by Acromine Disambiguator.

To explore more complex hybrid techniques, we experimented with chaining together alternative permutations of individual methods for each corpus, and found different optimal orderings in each case. The optimal orderings that we found for each corpus are shown in the bottom half of Table 7, where we illustrate how adding each method changes the performance of the previous combination. P-values are calculated based on the previous combination of results; we indicate with asterisks cases where adding a new method to the previous combination results in statistically significant performance improvements.

As shown in the results for clinical records, not all method additions result in performance increases. Although adding both SN and P2S to AAD+SS result in significantly significant performance increments, adding E2N offers little further improvement, and the addition of N2E is actually harmful, since it decreases precision quite considerably, despite increasing recall. Accordingly, the optimal method for clinical text uses only five of the six methods, and achieves a top performance of 0.7214 (compared to the dictionary lookup baseline of 0.6136 and the MetaMap baseline of 0.6682).

Table 7. Normalisation performance of hybrid methods (P = Precision; R = Recall; F = F-Score)

(a) Clinical records (11,167 mentions)

Methods	P	R	F	P-value
AAD	0.6775	0.6454	0.6610	–
AAD + N2E	0.6678	0.6547	0.6612	0.4340
AAD + SN	0.6767	0.6668	0.6717*	<0.0001
AAD + P2S	0.6564	0.6985	0.6768*	<0.0001
AAD + E2N	0.6772	0.6468	0.6617 *	<0.0001
AAD + SS	0.6689	0.7687	0.7154*	<0.0001
AAD + SS + SN	0.6672	0.7819	0.7200*	<0.0001
AAD + SS + SN + P2S	0.6653	0.7872	0.7212*	<0.0001
AAD + SS + SN + P2S + E2N	0.6647	0.7880	0.7214	0.5087
AAD + SS + SN + P2S + E2N + N2E	0.6458	0.8030	0.7159	–

(b) Biomedical abstracts (18,797 mentions)

Methods	P	R	F	P-value
AAD	0.9587	0.7163	0.8198	–
AAD + N2E	0.9515	0.7215	0.8207*	<0.0001
AAD + E2N	0.9572	0.7177	0.8204*	0.0011
AAD + SN	0.9542	0.7247	0.8238*	<0.0001
AAD + SS	0.9421	0.7996	0.8650*	<0.0001
AAD + P2S	0.9538	0.7262	0.8263*	<0.0001
AAD + P2S + N2E	0.9514	0.7315	0.8271*	<0.0001
AAD + P2S + N2E + E2N	0.9460	0.7389	0.8297*	<0.0001
AAD + P2S + N2E + E2N + SN	0.9397	0.7457	0.8315*	<0.0001
AAD + P2S + N2E + E2N + SN + SS	0.9313	0.8188	0.8714*	<0.0001

In contrast, for biomedical abstracts, all methods offer statistically significant incremental improvements as they are combined in the optimal order, with performance reaching 0.8714 when all methods are combined, compared to the dictionary-based baseline of 0.7741 and the MetaMap baseline of 0.8030. The interplay between individual methods is clearly evident in some cases. For example, when E2N is applied to biomedical abstracts after AAD+P2S+N2E, it improves the F-score by 0.026, compared to an improvement of only 0.004 when applied to the original concept mentions.

As expected, for both corpora, recall continues to increase as each new technique is added, demonstrating that each individual method can contribute additional successful mappings. Overall, recall increased consistently as more techniques were added into the hybrid method, achieving a top score of 0.8188 for biomedical abstracts (compared to 0.7251 for MetaMap). A top score of 0.7880 was achieved for clinical records, which cannot beat the recall of MetaMap (0.8781), but precision is much higher (0.6647 vs. 0.5394). Indeed, in both corpora, precision stays fairly consistent as new techniques are incrementally added, helping to show that, according to our aim, the hybrid methods retain their ability to generate variants that are mostly semantically consistent with the original mention.

The different optimal orderings of the methods for each corpus are likely to result from both the characteristics of the concept mentions, and the terminological resource that is the target of the normalisation. For example, the types of concept mentions that appear in biomedical abstracts can benefit more than clinical abstracts from the N2E and E2N methods. However, since the neoclassical-English dictionary lists base forms, the performance of these methods is improved by the prior application of the P2S, to find the base forms of nouns.

4.4 Discussion

Figure 5 shows some examples of concept mentions from both corpora that HYPHEN maps to the correct concept identifiers. Each line shows the specific normalisation method(s) that generated the successful matching variant, and the ID of the corresponding concept. Examples of cases requiring different individual methods to achieve the correct mapping are shown. To demonstrate the utility of our hybrid technique, we also show several cases in which multiple transformations must be undertaken to achieve the correct mapping.

A further important point to note from Figure 5 is that concept mentions in the overlap that appear in both corpora are mapped by HYPHEN to the correct identifiers in *both* SNOMED-CT and MeSH, which emphasises their flexibility.

An error analysis revealed a number of types of cases that are problematic for HYPHEN. Our requirement for semantic consistency between concept mentions and their generated variants means that all parts of the concept mention must be accounted for in the generated variant. Although this works in many cases, it is problematic for others, e.g., *cytotoxic oedema within cerebral white matter* in the abstract corpus cannot be mapped to *Brain Edema* (Doo1929) in MeSH, nor can *left lower lobe collapse* in the clinical record corpus be mapped to *atelectasis* (Cooo4144) in SNOMED-CT.

Particularly in MeSH, there may not be a concept identifier that corresponds to the exact concept described by the mention; in such cases, the gold standard

Figure 5. Examples of correct mappings achieved to MeSH (yellow), the SNOMED-CT (green), and both (overlap of boxes)

maps the mention to a more general concept, e.g. *renal dysfunction* is mapped to *kidney diseases* (Doo7674). Although N2E establishes an equivalence between *renal* and *kidney*, *dysfunction* is a more specific term than *diseases*, and so SS cannot help. In other cases, it appears that textual context is required to achieve the correct mapping. For example, whist HYPHEN maps the clinical record mention *aortic dissection* to the semantically equivalent [C0340643] *Dissection of aorta* in SNOMED-CT, the gold standard mapping is to the more specific [C0340647] *Distal aortic dissection*.

We also found that some gold standard normalisations in both evaluation corpora are actually incorrect, which will negatively impact our results. For example, HYPHEN maps the mention *difficult breathing* to *dyspnea* (Doo4417) in MeSH. However, the gold standard mapping is to the more general concept Do12120 (*respiration disorders*). In contrast, in the gold standard data for clinical records, a large number of concept mentions are designated “CUI-less”, i.e., they cannot be mapped to concepts in SNOMED-CT. However, HYPHEN assigned CUIs to many such cases, which we believe to be correct, e.g. *aphagia* (Co221470), *brain death* (Cooo6110), *choking* (Cooo8301), *instability* (C1444783) and *ecchymosis* (Coo13491).

4.5 Comparison with other methods

In this section, we compare the performance of HYPHEN with recent work. As mentioned above, both 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. The BioCreative V Chemical Disease Relation (CDR) Task used the same corpus of abstracts as our evaluation, but the disease named entity recognition task required disease mentions to be recognised *and* normalised. However, the best performing system (Lee et al. 2016) evaluates their dictionary-based normalisation technique in isolation on the gold standard annotation in the training part of the corpus, achieving an F-Score of 0.8415. When applied to the same portion of the corpus, HYPHEN achieves a higher F-score of 0.8649, which helps emphasise the benefits of applying a range of techniques for normalisation.

For clinical records, we can compare the performance of HYPHEN with that of PhenoNorm (Alnazzawi et al. 2016). When applied to the ShARe/CLEF corpus, PhenoNorm achieved a higher accuracy than HYPHEN (i.e. 0.83 vs. 0.72 accuracy), suggesting that the former is better suited to normalising concept mentions in clinical records, probably according to its use of string similarity methods. Although we have shown that these methods can result in the types of lexical-level errors that HYPHEN is designed to eliminate, they do allow more flexibility than our techniques, by allowing links to be established between mentions with differing word orders, word forms and numbers of words. As such, for example, PhenoNorm is able to map *stenosis in left anterior descending* to *left anterior descending coronary artery stenosis*.

5. Conclusions and future work

In this article, we have described our development of a new hybrid normalisation method, HYPHEN, which integrates a number of different normalisation techniques that account for a range of variations that can occur amongst concept mentions in different text types with widely varying characteristics. Even when used in isolation, some of these individual techniques can improve upon the normalisation performance achieved by the well-known MetaMap tool, for both biomedical abstracts and clinical records. However, when the different techniques are combined in an optimal order, their power becomes even greater. Generally, incremental increases in performance are observable as more techniques are integrated. Although there is some overhead in determining the optimal ordering of the methods on different corpora, we have shown that, once this has been car-

ried out, the hybrid methods are able to offer significant improvements in performance over MetaMap. Comparisons with other recent related work show that HYPHEN is particularly well suited to normalising phenotype concept mentions in biomedical text.

In its current state, HYPHEN achieves lower performance when applied to clinical records than abstracts, largely because of the unpredictability of mention variations in records. Although HYPHEN's performance on this text type is higher than that of MetaMap, we believe that integrating string similarity methods within HYPHEN could be helpful, as long as semantic equivalence can be maintained amongst generated variants. We also intend to investigate a number of other extensions, such as integrating other types of information from the terminological resources (e.g., hierarchical structure and concept definitions) to try to generate additional potential variants. Additionally, we plan to evaluate HYPHEN's ability to normalise a wider range of concept types, using different terminological resources, in a broader range of text types, which will include social media postings, since these constitute a further valuable source of medical information.

Acknowledgements

The work described in this article has been supported by the EPSRC and MRC (MMPathIC project, Grant. No. MR/N00583X/1), and the BBSRC (EMPATHY project, Grant No. BB/M006891/1).

References

- Alnazzawi, Noha, Paul Thompson, and Sophia Ananiadou. 2016. "Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource." *PLOS ONE* 11 (9): e0162287.
- Ananiadou, Sophia. 1994. "A Methodology for Automatic Term Recognition." In *Proceedings of the 15th Conference on Computational Linguistics—Volume 2*, 1034–1038, Kyoto, Japan. <https://doi.org/10.3115/991250.991317>
- Aronson, Alan R., and François-Michel Lang. 2010. "An Overview of Metamap: Historical Perspective and Recent Advances." *Journal of the American Medical Informatics Association* 17 (3): 229–236. <https://doi.org/10.1136/jamia.2009.002733>
- Bodenreider, O. 2004. "The Unified Medical Language System (Umls): Integrating Biomedical Terminology." *Nucleic Acids Research* 32: 267–270. <https://doi.org/10.1093/nar/gkh061>

- Bodnari, Andreea, Louise Deleger, Thomas Lavergne, Aurelie Neveol, and Pierre Zweigenbaum. 2013. "A Supervised Named-Entity Extraction System for Medical Text." In *Proceedings of the hARE/CLEF Evaluation Lab*, Valencia, Spain (<http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-BodnariEt2013.pdf>). Accessed 8 February 2018.
- Carroll, John, Rob Koeling, and Shivani Puri. 2012. "Lexical Acquisition for Clinical Text Mining Using Distributional Similarity." In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, 232–246, New Delhi, India. https://doi.org/10.1007/978-3-642-28601-8_20
- Cohen, William, Pradeep Ravikumar, and Stephen Fienberg. 2003. "A Comparison of String Metrics for Matching Names and Records." In *Proceedings of the KDD Workshop on Data Cleaning and Object Consolidation*, 73–78, Washington DC, USA.
- Collier, Nigel, Anika Oellrich, and Tudor Groza. 2015. "Concept Selection for Phenotypes and Diseases Using Learn to Rank." *Journal of Biomedical Semantics* 6 (1): 24. <https://doi.org/10.1186/s13326-015-0019-z>
- Dai, Manhong, Nigam H. Shah, Wei Xuan, Mark A. Musen, Stanley J. Watson, Brian D. Athey, and Fan Meng. 2008. "An Efficient Solution for Mapping Free Text to Ontology Terms." In *Proceedings of the AMIA Summit on Translational Bioinformatics*, San Francisco, USA (<https://knowledge.amia.org/amia-55142-tbi2008a-1.650887/t-002-1.985042/f-001-1.985043/a-041-1.985157/an-041-1.985158?qr=1>). Accessed 8 February 2018.
- Deléger, Louise, Fiammetta Namer, and Pierre Zweigenbaum. 2007. "Defining Medical Words: Transposing Morphosemantic Analysis from French to English." In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, 535–539, Brisbane, Australia.
- Doğan, Rezarta Islamaj, Robert Leaman, and Zhiyong Lu. 2014. "Ncbi Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization." *Journal of Biomedical Informatics* 47: 1–10. <https://doi.org/10.1016/j.jbi.2013.12.006>
- Dogan, Rezarta Islamaj, and Zhiyong Lu. 2012. "An Inference Method for Disease Name Normalization." In *Proceedings of the AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, 8–13, Arlington, USA.
- Donnelly, Kevin. 2006. "Snomed-Ct: The Advanced Terminology and Coding System for Ehealth." *Studies in Health Technology and Informatics* 121: 279.
- Duclos, C., A. Burgun, J.B. Lamy, P. Landais, J.M. Rodrigues, L. Soualmia, and P. Zweigenbaum. 2014. "Medical Vocabulary, Terminological Resources and Information Coding in the Health Domain." In *Medical Informatics, E-Health*, edited by A. Venot, A. Burgun and Quantin, 11–41. Paris, France: Springer. https://doi.org/10.1007/978-2-8178-0478-1_2
- Elhadad, Noémie, Sameer Pradhan, W.W. Chapman, Suresh Manandhar, and G.K. Savova. 2015. "Semeval-2015 Task 14: Analysis of Clinical Text." In *Proceedings of Workshop on Semantic Evaluation. Association for Computational Linguistics*, 303–310, Denver, USA.
- Fan, Jung-wei, Navdeep Sood, and Yang Huang. 2013. "Disorder Concept Identification from Clinical Notes an Experience with the Share/Clef 2013 Challenge." In *Proceedings of the ShARE/CLEF Evaluation Lab*, Valencia, Spain (<http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-FanEt2013.pdf>). Accessed 8 February 2018.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

- Fu, Xiao, and Sophia Ananiadou. 2014. "Improving the Extraction of Clinical Concepts from Clinical Records." In *Proceedings of BioTxtM14*, 47–53, Reykjavik, Iceland.
- Fu, Xiao, Riza Batista-Navarro, Rafal Rak, and Sophia Ananiadou. 2015. "Supporting the Annotation of Chronic Obstructive Pulmonary Disease (COPD) Phenotypes with Text Mining Workflows." *Journal of Biomedical Semantics* 6 (1): 8.
<https://doi.org/10.1186/s13326-015-0004-6>
- Fu, Xiao, R. T. B. Batista-Navarro, Rafal Rak, and Sophia Ananiadou. 2014. "A Strategy for Annotating Clinical Records with Phenotypic Information Relating to the Chronic Obstructive Pulmonary Disease." In *Proceedings of Phenotype Day ISMB*, 1–8, Boston, USA.
- Groza, Tudor, Sebastian Köhler, Dawid Moldenhauer, Nicole Vasilevsky, Gareth Baynam, Tomasz Zemojtel, Lynn Marie Schriml, Warren Alden Kibbe, Paul N. Schofield, and Tim Beck. 2015. "The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease." *The American Journal of Human Genetics* 97 (1):111–124.
<https://doi.org/10.1016/j.ajhg.2015.05.020>
- Habash, Nizar, and Bonnie Dorr. 2003. "Catvar: A Database of Categorical Variations for English." In *Proceedings of the MT Summit*, 17–23, New Orleans, US.
- Hamosh, Ada, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. 2005. "Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders." *Nucleic Acids Research* 33 (suppl 1):D514–D517.
<https://doi.org/10.1093/nar/gkio33>
- Han, MeiLan K., Alvar Agusti, Peter M. Calverley, Bartolome R. Celli, Gerard Criner, Jeffrey L. Curtis, Leonardo M. Fabbri, Jonathan G. Goldin, Paul W. Jones, and William MacNee. 2010. "Chronic Obstructive Pulmonary Disease Phenotypes: The Future of COPD." *American Journal of Respiratory and Critical Care Medicine* 182 (5): 598–604.
<https://doi.org/10.1164/rccm.200912-1843CC>
- Hersh, William R., and Robert A. Greenes. 1990. "Saphire – an Information Retrieval System Featuring Concept Matching, Automatic Indexing, Probabilistic Retrieval, and Hierarchical Relationships." *Computers and Biomedical Research* 23 (5): 410–425.
[https://doi.org/10.1016/0010-4809\(90\)90031-7](https://doi.org/10.1016/0010-4809(90)90031-7)
- Jaccard, Paul. 1912. "The Distribution of the Flora in the Alpine Zone." *New Phytologist* 11 (2): 37–50. <https://doi.org/10.1111/j.1469-8137.1912.tb05611.x>
- Jacquemin, Christian. 1999. "Syntagmatic and Paradigmatic Representations of Term Variation." In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 341–348, Maryland, USA.
<https://doi.org/10.3115/1034678.1034733>
- Jonquet, Clement, Nigam Shah, and Mark Musen. 2009. "The Open Biomedical Annotator." In *Proceedings of the AMIA summit on translational bioinformatics*, 56–60, San Francisco, USA.
- Kang, Ning, Rogier J. Barendse, Zubair Afzal, Bharat Singh, Martijn J. Schuemie, Erik M van Mulligen, and Jan A. Kors. 2010. "Erasmus MC Approaches to the I2b2 Challenge." In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, Boston, MA, USA
(http://biosemantics.org/downloads/i2b2_challenge.pdf). Accessed 15 February 2018.
- Kate, Rohit J. 2015. "Normalizing Clinical Terms Using Learned Edit Distance Patterns." *Journal of the American Medical Informatics Association* 23 (2): 380–386.

- Leaman, Robert, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. "Dnorm: Disease Name Normalization with Pairwise Learning to Rank." *Bioinformatics* 29 (22): 2909–2917. <https://doi.org/10.1093/bioinformatics/btt474>
- Leaman, Robert, Ritu Khare, and Zhiyong Lu. 2015. "Challenges in Clinical Natural Language Processing for Automated Disorder Normalization." *Journal of Biomedical Informatics* 57: 28–37. <https://doi.org/10.1016/j.jbi.2015.07.010>
- Leaman, Robert, Christopher Miller, and G. Gonzalez. 2009. "Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark." In *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, 82–89, Jeju Island, South Korea.
- Lee, Hsin-Chun, Yi-Yu Hsu, and Hung-Yu Kao. 2016. "Audis: An Automatic Crf-Enhanced Disease Normalization in Biomedical Text." *Database* 2016: bawo91.
- Li, Jiao, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wieggers, and Zhiyong Lu. 2016. "Biocreative V Cdr Task Corpus: A Resource for Chemical Disease Relation Extraction." *Database* 2016: bawo68.
- Maglott, Donna, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2011. "Entrez Gene: Gene-Centered Information at Ncbi." *Nucleic Acids Research* 39 (suppl 1): D52–D57. <https://doi.org/10.1093/nar/gkq1237>.
- Markó, Kornél, Stefan Schulz, Olena Medelyan, and Udo Hahn. 2005. "Bootstrapping Dictionaries for Cross-Language Information Retrieval." In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 528–535, Salvador, Brazil.
- Miyao, Yusuke, and Jun'ichi Tsujii. 2008. "Feature Forest Models for Probabilistic Hpsg Parsing." *Computational Linguistics* 34 (1): 35–80. <https://doi.org/10.1162/coli.2008.34.1.35>
- Namer, Fiammetta, and Robert Baud. 2005. "Predicting Lexical Relations between Biomedical Terms: Towards a Multilingual Morphosemantics-Based System." *Studies in Health Technology and Informatics* 116: 793–798.
- Névél, A., and P. Zweigenbaum. 2016. "Clinical Natural Language Processing in 2015: Leveraging the Variety of Texts of Clinical Interest." *IMIA Yearbook*: 234–239.
- Nunes, Tiago, David Campos, Sérgio Matos, and José Luís Oliveira. 2013. "Becas: Biomedical Concept Recognition Services and Visualization." *Bioinformatics* 29 (15): 1915–1916. <https://doi.org/10.1093/bioinformatics/btt317>
- Oellrich, Anika, Nigel Collier, Damian Smedley, and Tudor Groza. 2015. "Generation of Silver Standard Concept Annotations from Biomedical Texts with Special Relevance to Phenotypes." *PLOS ONE* 10 (1): e0116040. <https://doi.org/10.1371/journal.pone.0116040>
- Okazaki, N., S. Ananiadou, and J. Tsujii. 2010. "Building a High-Quality Sense Inventory for Improved Abbreviation Disambiguation." *Bioinformatics* 26 (9): 1246–1253. <https://doi.org/10.1093/bioinformatics/btq129>
- Patrick, Jon, Yefeng Wang, and Peter Budd. 2007. "An Automated System for Conversion of Clinical Notes into Snomed Clinical Terminology." In *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers*, 219–226, Ballarat, Australia.
- Pradhan, Sameer, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. "Semeval-2014 Task 7: Analysis of Clinical Text." In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 54–62, Dublin, Ireland. <https://doi.org/10.3115/v1/S14-2007>

- Pradhan, Sameer, Noémie Elhadad, Brett R. South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. 2015. "Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative." *Journal of the American Medical Informatics Association* 22 (1): 143–154.
- Rais, Meriem, and Natalia Grabar. 2015. "Discovering the Role of Morphology on the Understanding of Biomedical Terminology by Paramedical Students." In *Proceedings of the 26th Medical Informatics Europe Conference*, 30–34 Madrid, Spain.
- Ramanan, S. V., Shereen Broido, and P Senthil Nathan. 2013. "Performance of a Multi-Class Biomedical Tagger on Clinical Records." In *Proceedings of the ShARe/CLEF Evaluation Lab.*, Valencia, Spain (<http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-RamananEt2013.pdf>). Accessed 8 February 2018.
- Ruch, Patrick, Julien Gobeill, Christian Lovis, and Antoine Geissbühler. 2008. "Automatic Medical Encoding with Snomed Categories." *BMC Medical Informatics and Decision Making* 8 (1): S6. <https://doi.org/10.1186/1472-6947-8-S1-S6>
- Savova, Guergana K., James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. "Mayo Clinical Text Analysis and Knowledge Extraction System (Ctakes): Architecture, Component Evaluation and Applications." *Journal of the American Medical Informatics Association* 17 (5): 507–513. <https://doi.org/10.1136/jamia.2009.001560>
- Schriml, Lynn Marie, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2012. "Disease Ontology: A Backbone for Disease Semantic Integration." *Nucleic Acids Research* 40 (D1): D940–D946. <https://doi.org/10.1093/nar/gkr972>
- Suominen, Hanna, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, and Gareth J. F. Jones. 2013. "Overview of the Share/Clef Ehealth Evaluation Lab 2013." In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, 212–231, Valencia, Spain.
- Tanenblatt, Michael A., Anni Coden, and Igor L. Sominsky. 2010. "The Conceptmapper Approach to Named Entity Recognition." In *Proceedings of LREC*, 546–551 Valletta, Malta.
- Thompson, Paul, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, and Valeria Quochi. 2011. "The Biolexicon: A Large-Scale Terminological Resource for Biomedical Text Mining." *BMC Bioinformatics* 12 (1): 397. <https://doi.org/10.1186/1471-2105-12-397>
- Uzuner, Özlem, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. "2010 I2b2/Va Challenge on Concepts, Assertions, and Relations in Clinical Text." *Journal of the American Medical Informatics Association* 18 (5): 552–556. <https://doi.org/10.1136/amiajnl-2011-000203>
- Wang, Chunye, and Ramakrishna Akella. 2013. "Ucsc's System for Clef Ehealth 2013 Task 1." In *Proceedings of the ShARe/CLEF Evaluation Lab.*, Valencia, Spain (<http://ceur-ws.org/Vol-1179/CLEF2013wn-CLEFeHealth-WangEt2013.pdf>). Accessed 8 February 2018.
- Wang, Liqin, Bruce E. Bray, Jianlin Shi, Guilherme Del Fiol, and Peter J. Haug. 2016. "A Method for the Development of Disease-Specific Reference Standards Vocabularies from Textual Biomedical Literature Resources." *Artificial Intelligence in Medicine* 68: 47–57.
- Wulff, Henrik R. 2004. "The Language of Medicine." *Journal of the Royal Society of Medicine* 97 (4): 187–188. <https://doi.org/10.1258/jrsm.97.4.187>

Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. 2006. "Maxmatcher: Biological Concept Extraction Using Approximate Dictionary Lookup." In *Proceedings of PRICAI 2006: Trends in Artificial Intelligence*, 1145–1149, Guilin, China.
https://doi.org/10.1007/978-3-540-36668-3_150

Address for correspondence

Paul Thompson
University of Manchester
National Centre for Text Mining
Manchester Institute of Biotechnology
School of Computer Science
131 Princess Street
Manchester, M1 7DN
UK
paul.thompson@manchester.ac.uk

Author information

Sophia Ananiadou
National Centre for Text Mining
Manchester Institute of Biotechnology
School of Computer Science