PhenoCHF CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

PhenoCHF is an annotated corpus consisting of documents belonging to two different text types

- narrative reports from electronic health records (EHRs)
- literature articles

It is manually annotated by medical doctors with detailed information relating to mentions of phenotype concepts and disease-phenotype relations.

The documents in PhenoCHF focus on a specific medical condition, i.e., congestive heart failure (CHF). This focus is motivated by CHF's current standing as the world's most deadly disease. However, our experiments using the corpus have demonstrated that it can be used to develop systems that can recognise information relating to a wider range of diseases in a broader variety of text types than those included in PhenoCHF.

1.2 Representation of the corpus (flat files, database, markup)

The corpus consists of a set of plain text files (with *.txt* extensions), accompanied by files containing stand-off annotations (with *.a1* and *.a2* extensions)

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Institute of Biotechnology,131 Princess Street,
Manchester M1 7DN, IK
Affiliation: National Centre for Text Mining, School of Computer Science,
University of Manchester
Position: Professor
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the META-SHARE platform as an archive. Information about annotations is provided in separate files from the text that has been annotated.

The corpus consists of:

- A set of annotation files, containing the manually-added annotations associated with each text file.
- A set of text files corresponding to the literature articles only.

NOTE: The text files for the narrative EHR reports form part of the corpus de-identified clinical records released as part of the i2b2 2008 Obesity Challenge (NLP Dataset #2). The dataset must be obtained individually from Partners Healthcare by signing a Data Use Agreement (https://www.i2b2.org/NLP/DataSets/Main.php).

2.3 Copyright statement and information on IPR

The resource is licensed under a Creative Commons Attribution licence (CC-BY). If you use the resource, please attribute:

a) The National Centre for Text Mining (NaCTeM), who created the annotations. Please also cite the following article(s), depending on which types of annotations are used:

Entity Annotations

Alnazzawi, N., Thompson, P., Batista-Navarro, R. and Ananiadou, S. (2015). Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Medical Informatics and Decision Making*, 15(Suppl. 2): S3

Normalisation Annotations

Noha Alnazzawi, Paul Thompson and Sophia Ananiadou (2016). Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource. *PLOS ONE*.

Relation Annotations

Alnazzawi, N., Thompson, P. and Ananiadou, S.. (2014). Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi),* pp. 69-74, The full text literature articles in the PhenoCHF corpus are drawn from the PMC Open Access Subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/). These articles are protected by copyright, but are made available under a Creative Commons (https://creativecommons.org/about/license/) or similar licence that generally allows more liberal redistribution and reuse than a traditional copyrighted work. Please refer to the license of each article for specific licence terms.

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *PhenoCHF.* It contains the following sub-directories and files:

- *Articles* – directory containing the literature articles and associated annotations. The directory contains plain text files (.txt) and associated files containing stand-off annotations (with .a1 and .a2 extensions). These standoff files follow the format of the corpora created in the context of the BioNLP 2013 Shared Task, as described below.
  NarrativeEHR – directory containing annotation files (.a1 and .a2) for the narrative reports from Electronic Health Records (EHRs). The standoff files follow the format of the corpora created in the context of the BioNLP 2013 Shared Task, as described below. As mentioned above, the associated text for these EHR reports must be obtained separately from Partners Healthcare by signing a Data Use Agreement (https://www.i2b2.org/NLP/DataSets/Main.php).
- *phenoCHF_README.txt* – provides a range of information about the annotated corpus including the annotation scheme, composition of the corpus and the format of the annotation files.
- *phenoCHF_licence.txt* – Provides information about the licence applied to the corpus.

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

3.3 Resource size (nmb. of tokens, MB occupied on disk)

The corpus consists of 300 discharge summaries from EHRs and 10 full papers. It requires approximately 2.4 MB of disk space.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is English.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains full-text biomedical literature articles and narrative reports from electronic health records.

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Three levels of semantic information have been annotated in PhenoCHF:

- Named entities belonging to six different categories of medical significance.
- Normalisation annotations, which associate entities belonging to four of the annotated types with a Concept Unique Identifier (CUI), corresponding to a concept listed in the UMLS Metathesaurus (https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html)
- Three types of relations, which encode more complex information that is expressed about entities in text (e.g. whether they are negated, links between entities, etc.).

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

ENTITY ANNOTATIONS

Brief definitions of the six entity types annotated in PhenoCHF are provided in Table 1

Table 1. Entity types annotated in HIMERA

Entity Type	**Description**	**Examples**

Cause	Any medical problem that contributes to the occurrence of CHF	*chronic renal insufficiency, hypertension*
Risk Factor	A condition that increases the chance of a patient having the CHF disease	*obesity, type 2 diabetes, high cholesterol*
Sign & Symptom	Any observable manifestation of a disease which is experienced by a patient and reported to the physician	*productive cough, nausea, vomiting*
Non-traditional risk factor	Conditions associated with abnormalities in kidney functions that put the patient at higher risk of developing signs & symptoms and causes of CHF	*iron deficiency, anemia*
Organ	Any body part	*lungs, abdomen*
Chief Complaint	Mentions of CHF	*CHF, congestive heart failure*

All mentions of the concepts of the types shown in Table 1 were annotated in each document of PhenoCHF. The total counts of each type of entity annotated in each part of the corpus (i.e., narrative EHR reports and literature articles) are shown in Table 2.

Table 2. Statitics of Entity Mentions in PhenoCHF

Concept Type	**No of annotated mentions in narrative EHR reports**	**No of annotated mentions in literature articles**
Cause	1320	1107
Risk Factor	1335	408
Sign & Symptom	2449	304
Non-traditional risk factor	308	329
Organ	432	-

NORMALISATION ANNOTATIONS

All entity mentions belonging to four of the concept categories annotated during the entity mention annotation effort are associated with a unique concept listed in the UMLS Metathesaurus (https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html). This is a large terminological resource that contains entries for millions of biomedical and health related concepts. Each distinct concept in the UMLS Metathesaurus is assigned a unique identifier code called a Concept Unique Identifier (CUI).

The normalisation annotation in PhenoCHF involves a CUI being assigned by a medical doctor to all instances of the following types of entity mentions:

- Cause
- Risk Factor
- Non-tradtional Risk Factor
- Signs & Symptoms

The normalisation annotation in the PhenoCHF corpus provides the means to develop systems that can perform normalisation automatically.

Automatic normalisation can be important, given that each concept can be expressed in text in many different ways. Resources such as the UMLS Metathesaurus usually list some synonyms for each concept, i.e., different ways in which the concept could be expressed in text. However, there tend to be many more ways of mentioning a concept in text than those that correspond to the synonyms listed for the concept in terminological resources. Part of the problem is that such resources are usually manually curated. This means that is impossible to keep track of all possible ways in which a concept could be mentioned in text, especially according to the highly creative nature of language.

Automatic normalisation methods can help to identify appropriate CUIs for concept mentions that do not appear in the UMLS Metathesaurus.

Typically, synonyms of concepts listed in terminological resources tend to represent "standard" ways of referring to the concept (typically noun phrases), which often do not reflect how the concept is actually mentioned in text. Variation amongst mentions of concepts is particularly apparent in narrative EHR reports. For example, concepts may be mentioned in text as simple noun phrases (e.g. *progressive renal failure*), noun phrases followed by prepositional phrases (e.g., *increasing dyspnea on exertion*) and complete clauses or sentences (e.g., *jugular venous pressure is elevated*).

Table 3 illustrates some of the different types of variation that can occur amongst mentions of the same concept.

Table 3. Types of variation amongst concept mentions

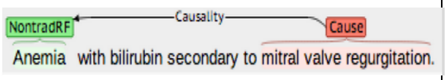
Type of Variability	EHR mentions	Article mentions
Synonymy	Sodium overload Drop in blood pressure	Hypernatremia Hypotension
Syntactic structure	Left ventricular is dilated Mild mitral calcification	Left ventricular dilatation Calcification of mitral valve
Word ordering	Cardiac output decreased	Decreased cardiac output
Morphological variation	Hyperkalemic	Hyperkalemia

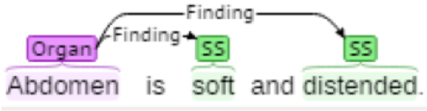
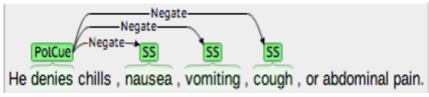
RELATION ANNOTATIONS

In the narrative EHR reports of PhenoCHF, three types of relations involving entity mentions have been annotated. Compared to entity mention annotation, which identifies mentions of concepts occurring in text, relation annotation encodes some of the more complex pieces of information expressed in text. Specifically, a sentence will often describe how different concepts are linked together in particular ways, or will provide a particular interpretation of an entity.

Table 4 provides details of the three types of relations that have been annotated in PhenoCHF (narrative EHR reports only). All annotated relationships occur within the scope of a single sentence. Two of these relationships involve pairs of entity mentions which, according to the information provided in the sentence, are associated with each other in specific ways. The semantic label assigned to the relation determines the nature of the association between the two entity mentions. The third relationship (Negate) is annotated when the context of the sentence alters the default interpretation of an entity (i.e., it becomes negated).

Table 4. Types of relations annotated in PhenoCHF

Relation Type	**Description**	**First entity type(s)**	**Second entity type(s)**	**Example**
Causality	The concept referred to by the first entity mention is responsible	Chief complaint Cause Risk Factor	Non-traditional Risk Factor Cause	 Anemia with bilirubin secondary to mitral valve regurgitation.

	for the concept refered to by the second entity mention	Non-traditional Risk factor		
Finding	The mentioned organ is associated with the manifestation or abnormal variation that is observed during the diagnosis process.	Organ	Sign & Symptom	 Abdomen is soft and distended.
Negate	A word or phrase denoting negation (a *polarity cue*) is annotated and linked to the mention of the condition that it negates	Polarity cue	Finding Cause Non-traditional risk factor	 He denies chills , nausea , vomiting , cough , or abdominal pain.

The two relationships between pairs of entity mentions are are based on relationships in the UMLS semantic network:

- **Causality** - based on the causes relation in the UMLS semantic network that holds between two diseases or a disease and a pathologic function
- **Finding** - based on the manifestation relation in the UMLS semantic network that holds between a sign or symptom and a body part or organ

All relationships of the types shown in Table 4 were annotated in all narrative EHR records. The total counts of each type of relationship annotated are shown in Table 5.

Table 5. Statitics of Event Mentions in PhenoCHF

Relation Type	**No of annotated relations in narrative EHR reports**
Causality	125
Finding	364
Negate	692

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

Annotations are encoded in the BioNLP Shared Task 2013 format (http://2013.bionlp-st.org/file-formats), with some custom additions to allow normalisation annotations to be encoded. Based on this format, there are two annotation files associated with each text file:

- **a1 files** - encode information about entity annotations, polarity cues and normalisation annotations
- **a2 files** - encode information about relation annotations.

A1 FILES

In a1 files, each line corresponds to an annotation. There are two formats of lines, depending on whether they encode an entity mention annotation or a normalisation. The format of each type of line is described below:

Entity Mention Annotations

Entity mention annotations encode the text spans corresponding to phenotype concept mentions (or polarity cues for Negate relations, see below), and assign a semantic label, according to the type of concept being mentioned.

A sample of lines encoding entity annotations is shown below:

```
T1     Cause 128 151   coronary artery disease
```

```
T5      NontradRF 285 291   anemia
T6      SignOrSymptom 393 412      shortness of breath
T2      RiskFactor 211 233  deep venous thrombosis
T8      SignOrSymptom 451 469      bilateral crackles
T9      Organ 440 445 Lungs
T10     RiskFactor 6272 6281;6282 6290
```

Each line that encodes an entity mention consists of the following information:

- **A unique id for the entity**. By convention, this starts with *T*, followed by a numerical value.
- **A TAB character.**
- The **concept type label** assigned to the annotation (or *PolCue* for words or phrases that denote negation, i.e., polarity cues). The labels corresponding to each concept type are shown in Table 4.
- The **character-based offsets** of the entity annotation in the corresponding text file. There are two formats for the offsets, depending on whether the annotated span consists of a single, continuous span or a *discontinuous* span, consisting of multiple, conncted spans. A discontinous span may occur, for example, when an entity mention is broken over two lines.
    - For continuous spans (as in the first 6 lines in the sample above), there are two offsets, corresponding to the start and end offsets of the span. The first offset is separated by a space from the entity type label, and there is a space between the start and end offsets.
    - For discontinuous spans (as in the final line of the sample above), there are two or more pairs of start and end offsets, each separated by a semi-colon. Each pair of offsets corresponds to a part of the complete annotated span.
- Another **TAB** character
- **The text covered by the annotated span** in the corresponding text file.

Table 6 provides the labels used for each concept type.

Table 6. Labels used in annotation files for each concept type or polarity cue

Concept type	**Label used in annotation file**
Cause	`Cause`
Risk Factor	`RiskFactor`
Sign & Symptom	`SignOrSymptom`
Non-traditional risk factor	`NonTradRF`
Organ	`Organ`

Polarity Cue	`PolCue`
Chief Complaint	`ChiefComplaint`

Normalisation Annotations

The normalisation annotations provide a mapping between each entity mention annotation and the identifier for a concept in the UMLS Metathesaurus (https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html) (i.e., a UMLS CUI).

A sample of lines encoding normalisation annotations is shown below:

```
#1      UMLS_CUI T1   C1956346
#2      UMLS_CUI T5   C0002871
#3      UMLS_CUI T6   C0013404
#4      UMLS_CUI T2   C0149871
#5      UMLS_CUI T8   C2071429
```

The format of these lines is as follows:

- **A unique numeric identifier for the normalisation annotation**. This is preceded by a hash character (#)
- A **TAB** character.
- **The string "UMLS_CUI"**
- The **identifier of the entity mention** annotation to which the UMLS CUI has been assigned.
- A **TAB** character.
- The **UMLS CUI** that represents the concept described by the entity mention.

A2 FILES

In a2 files, each line corresponds to a relation annotation.

Relation annotations have the following format:

```
R12     Causality Arg1:T18 Arg2:T17
R25     Finding Arg1:T64 Arg2:T66
R13     Negate Arg1:T41 Arg2:T37
```

Each line consists of:

- **A unique id for the relation annotation**. By convention, this starts with R, followed by a numerical value.
- A **TAB** character.
- The **Relation type label** assigned to the annotation. This one of: *Causality, Finding* or *Negate*.

- **Details of the two text spans that are linked in the relation**.
    - In the case of *Causality* and *Finding* relations, both text spans correspond to entity mentions.
    - In the case of *Negate* relations, the first of the text spans is a polarity cue for negation, while the second is an entity mention.

Each text span that is linked in a relation annotation is referred to as an *argument*. The first argument is denoted by the label *Arg1* and the second argument is denoted by the label *Arg2*. In each case, the argument label is followed by a colon, and then by the ID of the corresponding text span (which corresponds to one of the *T* annotations introduced above).

4.5 Intended application of the corpus

The composition and annotations in PhenoCHF are aimed at allowing the development of robust text mining (TM) systems that can extract comprehensive phenotypic information from multiple textual sources with differing characteristics. For example, narrative EHRs typically exhibit non-standard grammatical structure and high levels of lexical and semantic variability, coupled with many domain-specific abbreviations, complex sentences and spelling errors (around 10% of words).

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Entity Annotations

The entity annotations were undertaken by two medical doctors. The quality and consistency of the annotations were verified through the calculation of inter-annotator agreement (IAA). We calculated IAA in terms of F-Score, and found that high levels of agreement were acheived. We calcluated both exact span matches, where the start and end of the annotated text spans chosen by both annotators must match exactly, and relaxed span matches, where it is sufficient for the annotated text spans to include some common parts. The IAA statistics, in terms of F-score, are shown in Table 5.

Agreement Type	**Narrative EHRs**	**Literature articles**
Exact Match	0.82	0.69
Relaxed Match	0.92	0.77

Relation Annotations

The relationship annotations were undertaken by two medical doctors. The quality and consistency of the annotations were verified through the calculation of inter-annotator agreement (IAA). We calculated IAA in terms of F-Score, and found that high levels of agreement were acheived (i.e., a macro-averaged F-Score of 0.91).

5 RELEVANT REFERENCES AND OTHER INFORMATION

The annotations in the PhenoCHF corpus are introduced in the following articles:

Entity Annotations

Alnazzawi, N., Thompson, P., Batista-Navarro, R. and Ananiadou, S.. (2015). Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Medical Informatics and Decision Making*, 15(Suppl. 2): S3

Normalisation Annotations

Noha Alnazzawi, Paul Thompson and Sophia Ananiadou (2016). Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource. *PLOS ONE*.

Relation Annotations

Alnazzawi, N., Thompson, P. and Ananiadou, S.. (2014). Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi),* pp. 69-74,