Deliverable D10.2

Project Title:	Building data bridges between biological and medical infrastructures in Europe	
Project Acronym:	BioMedBridges	
Grant agreement no.:	284209	
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"	
Deliverable title:	A prototype linking ICD10/SNOMED CT concepts to Ensembl gene identifiers	
WP No.	10	
Lead Beneficiary:	5: UCPH	
WP Title	Integrating disease related data and terminology from samples of different types	
Contractual delivery date:	30 June 2014	
Actual delivery date:	18 August 2014	
WP leader:	Alvis Brazma	1: EMBL
Partner(s) contributing to this deliverable:	1: EMBL, 3: KI, 15: UCPH	

Authors: Albert Pallejà Caro, Sune Frankild, David Westergaard, Pope Moseley & Søren Brunak

Contents

1 Executive summary

The main aim in of this deliverable is to produce a mapping between disease terminologies used to characterize patients' conditions on one side and biobank samples and human gene identifies on the other. This mapping then makes it possible to carry out systems biology analyses with the aim of finding the underlying causes of disease and, for example, also the molecular links between diseases and their comorbidities.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Linking disease-related data to molecular information: terminology	X	
2	Linking disease-related data to molecular information: data		X

3 Detailed report on the deliverable

3.1 Background

The healthcare sector extensively uses the WHO International Classification of Diseases (ICD)[1] to codify diseases when they diagnose patients in electronic and non-electronic patient health records. ICD is a hierarchical classification of diseases and symptoms divided into 22 anatomical or functional chapters. The main aim of this deliverable is to produce a mapping between disease terminologies used to characterize patients' conditions on one side and biobank samples and human gene identifies on the other. This mapping then makes it possible to carry out systems biology analyses with the aim of finding

[1] http://apps.who.int/classifications/icd10/browse/2010/en

the underlying causes of disease and, for example, also the molecular links between diseases and their comorbidities.

Specifically, we needed a mapping between ICD10 codes and genes and proteins (e.g. Ensembl IDs) associated with these diseases. To map the ICD10 codes to Ensembl proteins, we created an intermediate mapping between ICD10 and the Disease Ontology[2] (DO). To our knowledge, no mapping between the ICD10 and DO terms is publicly available. The DO is a standardized hierarchical ontology for human diseases. It provides the biomedical community with consistent descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts. This ontology semantically integrates disease and medical vocabularies by cross mapping of DO terms to MeSH, ICD version 9, NCI's thesaurus, SNOMED and OMIM identifiers. At the same time this ontology is much more effective in the context of text mining than ICD10. For the relationship between Disease Ontology terms and genes, we have extensive knowledge from both text mining (described further below), Uniprot[3] and the Genetic Home Reference (GHR)[4].

Both ICD10 and DO dictionaries are phenotype-oriented, which makes it possible for us to make a reliable mapping between them. We included all the ICD10 codes at level 3 contained in the chapters specified in Table 1. These are the chapters that we focus on for supporting subsequent studies to find the molecular link between comorbid diseases in the prototype. The lower the level in the classification, the more specific the ICD terms become. ICD10 level 3 codes are specified by a chapter letter and two numbers (e. g. obesity is E66). Lower levels were mapped to level 3 (e.g. E66.0-Obesity due to excess calories, E66.1-Drug-induced obesity and so on were mapped to E66). A more fine-grained mapping in a major ontology-wide attempt to map codes and genes will not be meaningful given the present status of the knowledge of the function of human genes, including their possible pleiotropic aspects.

[2] http://disease-ontology.org/
[3] http://www.uniprot.org/diseases/
[4] http://ghr.nlm.nih.gov/

Table 1 ICD 10 Chapters and number of codes mapped to DO terms

Chapters	ICD Blocks	ICD10 level 3 codes mapped to DO terms	DO terms used
II, Neoplasms	C00-D48	137	93
III, Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	33	17
IV, Endocrine, nutritional and metabolic diseases	E00-E90	73	38
V, Mental and behavioral disorders	F00-F99	78	27
VI, Diseases of the nervous system	G00-G99	68	32
VII, Diseases of the eye and adnexa	H00-H59	43	19
VIII, Diseases of the ear and mastoid process	H60-H95	24	4
IX, Diseases of the circulatory system	I00-I99	77	42
X, Diseases of the respiratory system	J00-J99	64	25
XI, Diseases of the digestive system	K00-K93	71	40
XII, Diseases of the skin and subcutaneous tissue	L00-L99	72	45
XIII, Diseases of the musculoskeletal system and connective tissue	M00-M99	70	30
XIV, Diseases of the genitourinary system	N00-N99	82	41
XV, Pregnancy, childbirth and the puperperium	O00-O99	67	18
XVII, Congenital malformations, deformations and chromosomal abnormalities	Q00-Q99	87	46

3.2 Mapping

We were able to manually map 1049 ICD10 level 3 codes to 489 DO terms (Table 2; the mapping is available in Supplement 2). The following two rules were applied when mapping ICD10 codes to equivalent DO terms: 1) if there was a direct or a meaningful match between an ICD10 code description and a DO term or one of its synonyms, we assigned the ICD10 code to that DO term; 2) if there was not a meaningful match between an ICD10 code and a DO term, we went up the DO hierarchy until we found a parent DO term that could include the ICD10 code and all its complexity (all its lower levels). Consequently, the mapping is in the direction: ICD10 to DO. The reciprocal mapping is not always true, only with the one-to-one mappings.

Table 2 Top 30 most gene rich ICD10 codes. The full list of ICD10 codes and number of gene-disease links is available in a tab-separated-value file (Supplement 1)

ICD10 level 3 code	Disease-gene links	ICD10 level 3 description
K87	1471	Disorders of gallbladder, biliary tract and pancreas in diseases classified elsewhere
Q38	1471	Other congenital malformations of tongue, mouth and pharynx
N99	1471	Postprocedural disorders of genitourinary system, not elsewhere classified
M79	1471	Other soft tissue disorders, not elsewhere classified
G83	736	Other paralytic syndromes
D38	590	Neoplasm of uncertain or unknown behaviour of middle ear and respiratory and intrathoracic organs
D39	590	Neoplasm of uncertain or unknown behaviour of female genital organs
D37	590	Neoplasm of uncertain or unknown behaviour of oral cavity and digestive organs
D48	590	Neoplasm of uncertain or unknown behaviour of other and unspecified sites

Table 2 (continued)

D43	590	Neoplasm of uncertain or unknown behaviour of brain and central nervous system
D42	590	Neoplasm of uncertain or unknown behaviour of meninges
D41	590	Neoplasm of uncertain or unknown behaviour of urinary organs
D40	590	Neoplasm of uncertain or unknown behaviour of male genital organs
D47	590	Other neoplasms of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue
D46	590	Myelodysplastic syndromes
D45	590	Polycythaemia vera
D44	590	Neoplasm of uncertain or unknown behaviour of endocrine glands
C97	582	Malignant neoplasms of independent (primary) multiple sites
C80	582	Malignant neoplasm, without specification of site
G23	337	Other degenerative diseases of basal ganglia
G08	337	Intracranial and intraspinal phlebitis and thrombophlebitis
G09	337	Sequelae of inflammatory diseases of central nervous system
G06	337	Intracranial and intraspinal abscess and granuloma
G07	337	Intracranial and intraspinal abscess and granuloma in diseases classified elsewhere
G97	337	Postprocedural disorders of nervous system, not elsewhere classified
G99	337	Other disorders of nervous system in diseases classified elsewhere
G82	337	Paraplegia and tetraplegia
G96	337	Other disorders of central nervous system
G98	337	Other disorders of nervous system, not elsewhere classified
E80	292	Disorders of porphyrin and bilirubin metabolism

As mentioned, the approach involves a text mining step working from millions of published papers. Abstracts of studies of genetic or functional causation co-mention both diseases, disease terms and genes found to be associated with risk. Named Entity Recognition (NER) systems are able to mine out this information on a large scale if one has a functional dictionary of disease names. However, ICD10 comes with a limited set of disease names and synonyms which are not suitable for text mining as they are meant for healthcare billing etc. and not text mining. The Disease Ontology (DO) is much more suitable and contains a rich set of disease names that match the way diseases are written in text in the published literature. We have therefore used the Disease Ontology for matching disease names/terms and the STRING dictionary for matching genes and proteins. The mapping between ICD10 and DO allows us thus to convert the text mining output to links between ICD10 disease codes and Ensembl protein IDs.

We have previously shown that a simple statistical scoring of the co-mentioning of proteins and drugs is an efficient estimator of molecular interaction networks[1]. The purpose of the text-mining pipeline here is to output scores between pairs of identifiers; in this case, scores between DO and Ensembl protein identifiers. In the quantification, the score S is a combination of a count C and an enrichment factor E. The count C is a weighted sum of unique co-mentioning pairs on the level of the document, paragraphs and sentences with the following weights: 1.0, 2.0 and 0.2. The enrichment factor E is the ratio of the observed count C to the expected count Cexp assuming a random co-mentioning frequency. However, since the literature constantly grows, scores depend in practice on the size of the corpus. To normalize for this, we transformed count-based scores S into the final Z-scores, Z. We then estimated the mean (μ) and standard deviation (**σ**) of the random S-score distribution and computed Z = (S-**μ**)/**σ**. From a benchmarking (data not included here), we found that the method has a precision of approximately 90% for pairs with Z-scores larger than 4.0. Thus, we expect that 9 out of 10 scores reported herein should constitute associations with valid text-mining evidence. The raw output from the text-mining pipeline contains information on 984 ICD10 codes and 14,817 proteins. However, when restricting this set to pairs with a score above 4.0 these numbers drop to 881 ICD10 codes and 1,832 unique genes, respectively. The

Uniprot (GHR) database contains information on 368 (431) ICD10 codes, linked to 1651 (965) unique proteins. Combining all of the data yields information on 883 ICD10 codes, for which there is at least one protein associated, 2,982 unique proteins and 59,828 ICD10-protein associations.

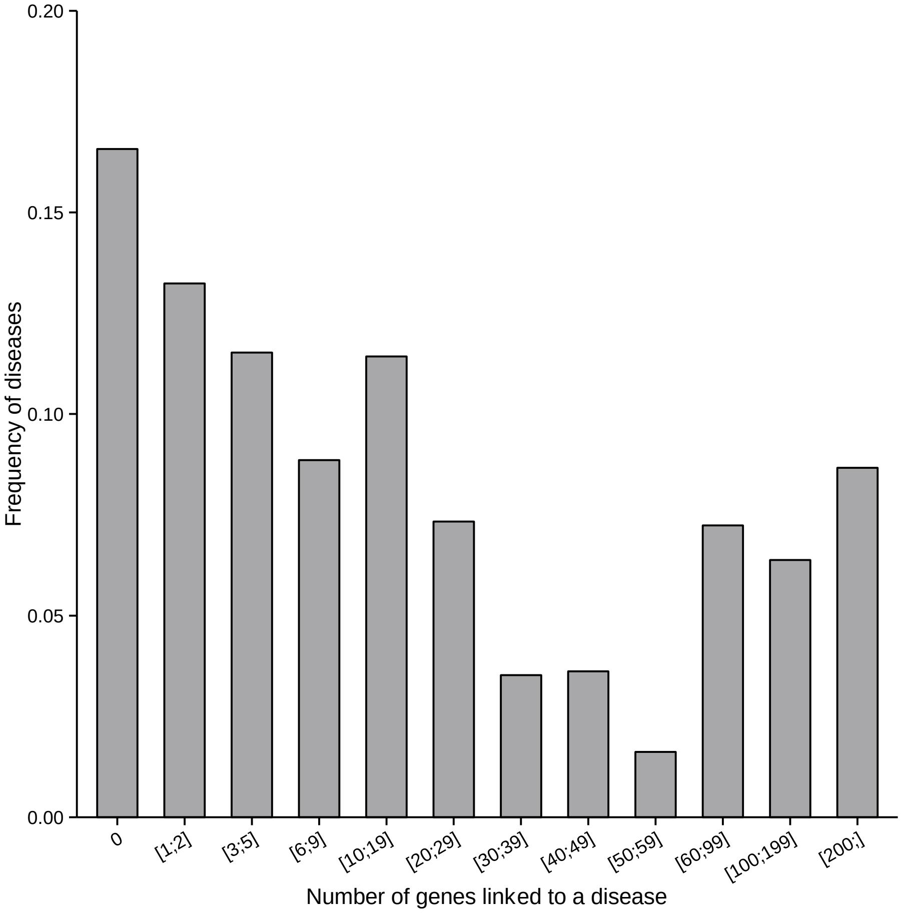

Figure 1 Histogram of the number of genes linked to a disease. The bins are not equidistant but reflect that the number of gene associations per disease code span two orders of magnitude

The distribution of the number of genes linked to ICD10 codes is shown in Figure 1. The most common number of (non-zero) genes per ICD10 disease is 1-2 genes. For 167 diseases (15.9%) there were no gene associations. The median number of gene associations is 11 (mean: 56) and one may notice that the distribution is slightly bimodal reflecting two major groups of disease. The lower part of the distribution reflects diseases with 0-50 genes and the upper reflects disease with 60+ genes. This class of ICD10 disease codes reflects very broad groups that are mapped onto equivalent broad groups in the Disease Ontology. The diseases linked to most genes (1471 associations) are thus:

K87: "*Disorders of gallbladder, biliary tract and pancreas in diseases classified elsewhere*"

Q38: "*Other congenital malformations of tongue, mouth and pharynx*",

N99: “*Postprocedural disorders of genitourinary system, not elsewhere classified*”

M79: “*Other soft tissue disorders, not elsewhere classified*”. All four are very broad disease categories.

Looking at the top 30 most gene rich ICD10 disease, the number of proteins ranges from 1471 to 282.

The number of disease–gene relations over ICD10 chapters is shown in Figure 2. It is apparent that the medical research field is dominated by cancer research, studies covering malformations, and endocrine, nervous, circular and mental disorders.

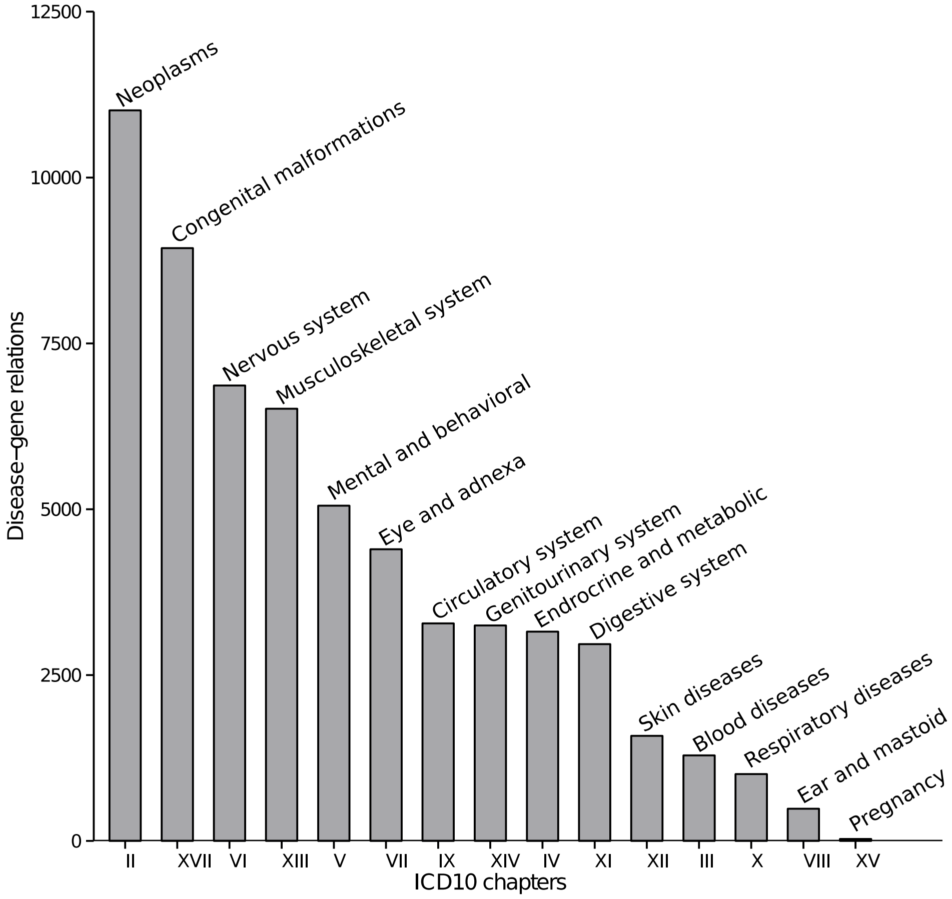

Figure 2 The number of disease-gene relations in ICD10 chapters. It is apparent that the medical literature is dominated by cancer research, studies covering malformations, and nervous, musculoskeletal and mental disorders

3.3 Further work

The deliverable presented in this paper relies heavily on two aspects:

1) The mapping between ICD10 and DO terms and
2) The mapping between DO terms and Ensembl protein identifiers.

The Disease Ontology is an ongoing effort, and as the project matures, terms will be added and others made obsolete. To ensure a consistent high quality of the data presented here, there is a need to monitor and react to updates made in the Disease Ontology. If a term is made obsolete, it is important to remove this from the mapping and, likewise, if a term is added which is a better descriptor of an ICD10 code, it is also important to add this to the mapping. Further to this, as the biomedical literature grows, it is important to continually monitor the text mining pipeline, Uniprot and the Genetic Home Reference.

4 References

[1] Franceschini A1, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ., “STRING v9.1: protein-protein interaction networks, with increased coverage and integration”, Nucleic Acids Res. 2013 Jan; 41 (Database issue):D808-15, http://dx.doi.org/10.1093/nar/gks1094

5 Supplementary information

Supplement 1: Full list of ICD10 codes and number of gene-disease links (tab-separated-value file): table_icd10_gene_count_descr.tsv

Supplement 2: Mapping (tab-separated-value file): *ICD10_to_doid.tsv*

6 Delivery and schedule

The delivery is delayed: ☑ Yes ☐ No

7 Adjustments made

No adjustments were made.

8 Background information

This deliverable relates to WP 10; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 10 Title: Integrating disease related data and terminology from samples of different types
Lead: Alvis Brazma (EMBL)
Participants: EMBL, KI, UCPH

This work package will demonstrate the feasibility and provide a prototype for linking disease to molecular information on two levels – terminology and data. It has two tasks respectively – first to link ICD10 terms to gens and protein complexes, and second to link data in selected BBMRI biobanks to samples at the EBI BioSample Database supported by ELIXIR.

Both tasks are related, as to link biobanks to ELIXIR databases, the respective terminologies have to be mapped. For completing the first task we will create a prototype for an interoperable scheme linking IICD10 to genes and protein complexes. This will enable linking biobank and other phenotypic healthcare sector data on individuals to individual genotypes. Interoperability schemes of this kind will be essential for linking BBMRI and ECRIN data to the molecular level and for linking biobank and other phenotypic healthcare sector data on individuals to individual genotypes.

To accomplish the second task, we will select a small number of biobanks participating in the BBMR infrastructure, with best advanced sample representation in databases. We will develop a model for linking sample objects to the respective objects at the BioSample database developed at EBI. The final deliverable in this work-package will be implementation of these links, allowing user to navigate from the selected biobanks to the EBI BioSample database and vice versa.

The work package will build upon standards developed in WP3, will utilize infrastructure build in WP4 and will benefit from data security framework developed in WP5.

Work package number	WP10	**Start date or starting event:**	month 13
Work package title	Integrating disease related data and terminology from samples of different types		
Activity Type	RTD		

Participant number	1:EMBL	3:KI	15: UCPH
Person-months per participant	26	20	31

Objectives

1. Linking disease-related data to molecular information: terminology
2. Linking disease-related data to molecular information: data.

Description of work and role of participants

Task 1. Mapping between sample information representation in a selected subset of resources. We will map data elements describing sample information in selected biobank databases that participate in the BBMRI federated infrastructure and the BioSample Database at EMBL-EBI. We will work jointly with WP3 to generalize the defined mappings and to develop the minimum standard that would enable to exchange this information.

Task 2. The aim is to create a prototype which maps existing healthcare sector terminologies and their phenotypic descriptions to existing repositories linking diseases, symptoms and genes. More generally the aim is to embed the prototype efficiently into exiting computational linguistics, bioinformatics and clinical environments alike. In particular we will link phenotypic terminology in healthcare sector data to genes we will create a prototype for an interoperable scheme linking ICD9 an ICD10 to gene identifiers in ELIXR database concentrating on genes that have been associated with specific diseases, symptoms and tissues. The prototype will focus specifically on the generic ICD concepts and their mapping to relevant genetic information, and will not include efforts which aim for assigning codes and terminology to free text in healthcare sector data as it usually is done by text mining. The work on the prototype will also include language interoperability aspects on the terminology side, while considering only gene names as they are used in English.

Task 3. The objective of this task is to demonstrate the interoperability of the sample representation at the EMBL-EBI’s BioSample database, with the biobank databases participating in the federated biobank information infrastructure created under BBMRI. This will also build upon the work done on terminology and identifier mapping. Working together with WP4 and WP5, we will implement a pilot for federated queries and secure links between a limited number of selected resources in BBMRI and the BioSample Database at EMBL-EBI (ELIXIR). A query system will be developed that will allow sample property based queries across these resources.