Using Dependence Relations in MeSH as a Framework for the Analysis of Disease Information in Medline

Lowell Vizenor
Medical Ontology Research
U. S. National Library of Medicine
Bethesda, MD 20894 - USA
vizenorl@mail.nih.gov

Olivier Bodenreider
Medical Ontology Research
U. S. National Library of Medicine
Bethesda, MD 20894 - USA
olivier@nlm.nih.gov

Abstract

Motivation: Many terminologies such as MeSH have a hierarchical structure, but no trans-ontological relations. Such relations, however, are useful for characterizing the relations identified in data sets. **Methods**: We identified trans-ontological relations in MeSH (between diseases and other categories) based on formal ontological principles and compared them to co-occurrence data in MEDLINE. **Results**: Dependence relations identified between a disease and other categories generally correspond to the highest proportion of relations between this disease and any other category under investigation. The other relations observed between this disease and other categories correspond mostly to contingent relations.

1 Introduction

There exist many kinds of relationships between biomedical entities. For example, a viral meningitis is *located in* the meninges, it is *caused by* some virus, and it can be *treated by* some antiviral drugs. Such relations are recorded explicitly as symbolic relations in biomedical knowledge bases and, to a lesser degree, in terminological resources such as SNOMED CT and MeSH. Moreover, the association between indexing terms (i.e., term co-occurrence) in the citations from a bibliographic database such as MEDLINE also represents statistical relations among these indexing terms. For example, the MeSH term *Viral meningitis* co-occurs frequently with MeSH terms for various virus species, including *Enterovirus B, Human* and *Herpervirus 2, Human*. One major difference between symbolic and statistical relations is that, whereas the nature of the symbolic relations is explicit (e.g., *location of*), the nature of the statistical relations is implicit. However, the frequency of co-occurrence can be analyzed to assess the salience of the association.

Formal ontology provides another perspective on relations, distinguishing between two major kinds of symbolic relations: dependence relations (inherent to the nature of related entities) and contingent relations. These distinctions are presented in detail in section 2.

The objective of this study is to analyze dependence relations in MeSH and to compare them to statistical relations obtained from co-occurrence data. We restrict our analysis to the relations between disease categories and other categories of biomedical interest.

Our hypothesis is that systematic associations will be found between diseases and the types of entities on which they are dependent, namely between diseases classified by location and their corresponding anatomical sites and between diseases classified by etiology and their corresponding causes or agents. In practice, for a given disease, the largest proportion of relations to another category should be to a category on which this disease is dependent, and this systematically for each disease. In contrast, we expect to find a smaller proportion of relations between diseases and other categories of biomedical interest, corresponding to contingent relations.

Besides clarifying the link between dependence relations and co-occurrence, this paper seeks to identify associative relations in MeSH (i.e., relations across trees), which, we expect, will support information retrieval and semantic mining applications.

2 Background

2.1 Relations in Biomedicine

Broadly speaking, there are two types of meaningful relations that hold between biomedical classes. First, there are those relations in which every instance of a class is related to some instance of another class. For example, *every* instance of protein synthesis is (necessarily) related to *some* strand of mRNA.

Second, there are those relations in which only some instances of a class are related to some instances of another class. For example, *some* aspirin tablets are related to *some* headaches. Certainly, the relation between aspirin and headaches is important (the former alleviates the latter), but merely contingent.

Here, though, we will focus exclusively on one type of necessity relation, namely, the relation of ontological dependence. Briefly, an entity *A* is ontologically dependent on *B* if and only if *A* exists then *B* exists. Moreover, we are interested in dependence relations between what we refer to as biomedical continuants and biomedical processes. But before we can state these relations more precisely, we need to introduce some ontological distinctions.

2.2 Formal-Ontological Distinctions

All real world entities in the biomedical domain fall into one of two exclusive categories of continuant and process. Think of the difference between a human being and the event of losing weight. Informally, what changes (the human) is the continuant and the change itself (the weight-loss event) is the process. More precisely, continuants are entities which continue to exist through time; they preserve their identity from one moment to the next even while undergoing a variety of different sorts of changes (Smith and Grenon 2004). Examples include molecules, cells, anatomical structures, and organisms.

Processes differ from continuants in several important respects. Most importantly, though, whereas continuants exist fully at a given time—i.e. all their parts are present at a time—processes never exist in full at a time; instead, they unfold through successive phases. Processes have a beginning, middle and end (Smith and Grenon 2004). Examples include cell division, ion transport, protein synthesis and respiration.

In addition to the distinction between continuants and processes, there is a distinction between independent and dependent continuants. *Independent continuants* (also sometimes called substances) are entities such as organisms, organs, cells and genes which do not require the existence of any other entity in order to exist. *Dependent continuants* (sometimes called accidents) are entities such as dispositions, functions, properties, qualities, roles and states. A dependent continuant is such as to be fully present at a given time, but nevertheless requires the existence of some independent continuant in order to exist.

Likewise, every process depends for its existence on some (independent) continuant. So, for example, *every* instance of photosynthesis (process) depends on *some* instance of chlorophyll (continuant).

The ontological categories of continuant and process provide a basis for the distinction between *intra-ontological relations* and *trans-ontological relations* (Grenon, Smith and Goldberg 2004; Smith and Grenon 2004; Rosse et al. 2005). Examples of intra-ontological relations are *subsumption* (the relation of one class being wholly included in another), *instantiation* (the relation between an individual [or instance] and a universal [or class]) and *meronomy* (the part-whole relation). These relations are always restricted to a given ontological category. A sound ontological principle for the construction of hierarchies is that all the classes employed should belong to one and only one of these categories. So, no continuant has a process as a part and vice versa. Similarly, no continuant is a subclass of a process and conversely. Trans-ontological relations are those sorts of relations that transcend the ontological divide between the formal ontological categories of continuant and process and are just those ontological dependence relations that hold between continuants and processes.

2.3 Dependence Relations between Biomedical Continuants and Processes

Think of the relation between the process of cell transport and the cell that participates in this process. Without the cell there is no cell transport. What is more, every instance of the processual class *cell transport* is ontologically dependent on some instance of the continuant class *cell*. *Participation* is the relation that holds between any continuant that is involved in some form or another in a process. Examples of participation include (Smith et al. 2005):

death has_participant organism

breathing has_participant thorax.

In what follows, we focus on the (implicit) participation relations in MeSH between processual diseases classified in terms of the anatomical structure or bodily system in which they are located.

(Smith and Grenon 2004) note that special types of participation relations can be distinguished according to whether a continuant is agent or patient in a process. In the case of *disease D has_participant anatomical structure S*, the continuant *S* is a patient in the process *D*. In other words, *S passively participates in D*.

But there are also cases where the continuant is an agent in the process. That is, *C actively participates in P*. For example, a pathogen actively participates in (i.e. is the cause of) a pathological process. We refer to relations of active participation as *has_agent*.

These relations are defined more formally in (Smith et al. 2005) as follows:

> *p* **has_participant** *c* **at** *t* – a primitive relation between a process (*p*), continuant (*c*), and a time (*t*)
>
> *p* **has_agent** *c* **at** *t* – a primitive relation between a process (*p*), a continuant (*c*) and a time (*t*) at which the continuant is causally active in the process.

Note that these relations are defined at the level of *instances*. Because terminologies such as MeSH contain essentially names for biomedical classes, we are primarily interested in relations between *classes*. (Smith et al. 2005) define the class-level counterparts of these relations as follows:

> *P has_participant C* = [definition] for all *p*, if *Pp* then there is some *c*, *t* such that *Cct* and *p* **has_participant** *c* **at** *t*
>
> *P has_agent C* = [definition] for all *p*, if *Pp* then there is some *c*, *t* such that *Cct* and *p* **has_agent** *c* **at** *t*

(where *Pp* reads *p* is an instance of processual class *P* and *Cct* reads *c* is an instance of continuant class *C* at time *t*).

2.4 Statistical Relations

In probability theory, two events E_1 and E_2 are independent when the probability of occurrence of the two events simultaneously, $P(E_1 \cap E_2)$, is not greater than the product of the probabilities of occurrence for each event, $P(E_1) \cdot P(E_2)$. Conversely, when $P(E_1 \cap E_2) > P(E_1) \cdot P(E_2)$, E_1 and E_2 are not independent. What we are interested in identifying here are pairs of "non-independent" MeSH terms, whose frequency of co-occurrence (i.e., simultaneous presence as indexing terms in a MEDLINE document) is higher than would be expected if the two terms had been used independently by the indexers.

3 Materials

3.1 MeSH

The Medical Subject Headings (MeSH®) is the controlled vocabulary used by the National Library of Medicine (NLM) for indexing articles from 4,800 of the world's leading biomedical journals for the MEDLINE/PubMED® database. The MeSH thesaurus is organized in 16 tree-like hierarchies. Examples of top-level categories in MeSH include *Anatomy*, *Organisms* and *Diseases*. Hierarchical relations among MeSH descriptors are indicated by tree numbers assigned to the descriptors. For example, the tree number **C05**.550 indicates that *Joint Diseases* is a descendant of *Musculoskeletal Diseases* [C05] in the C tree (*Diseases*). The 2004 version of MeSH is used in this study and comprises 22,658 descriptors.

3.2 MEDLINE

MEDLINE® is NLM's premier bibliographic database that contains approximately 13 million references to journal articles in life sciences with a concentration on biomedicine. MEDLINE citations are indexed manually using MeSH as a controlled vocabulary. Descriptors denoting the central topic of an article are indicated with a star. In this study, the co-occurrence of MeSH descriptors is restricted to co-occurrence between starred descriptors. In order to focus on salient associations, only those pairs of MeSH descriptors co-occurring at least 10 times in our collection are considered. The set of MEDLINE citations used in this study contains 385,642 citations and consists of the citations entered in MEDLINE and completed (indexed) between December 2003 and November 2004. Of these, we discarded 151 citations with no descriptor indicating a central topic, retaining a total of 385,491 citations. 20,085 distinct MeSH descriptors were used to index these citations, resulting in 1,378,027 citation-descriptor associations.

4 Methods

4.1 Identifying Dependence Relations

In this part of our study, we identify dependence relations across hierarchies manually, by examining the top-level categories in MeSH in the light

of both the formal ontological principles presented earlier and domain knowledge. We focus on relations between diseases and other categories, distinguishing between dependence and contingent relations. Moreover, we restrict our study to those diseases that are best understood as pathological processes. Note that not all diseases are processes. Think, for example, of velocardiofacial syndrome.

The first step, then, is to identify those MeSH disease descriptors that can be interpreted as a pathological process. For this reason we omit *Neoplasms* [C04], which is a term that refers to pathological continuants (i.e, tumors). The second step involves identifying the biological continuants that participate (actively or passively) in these pathological processes. Once we identify the dependence relations that exist between diseases and other categories we contrast these relations with contingent relations.

4.2 Identifying Statistical Relations

For a given pair of MeSH terms (*A*,*B*), information about their association in documents can be summarized in a two-way contingency table and analyzed statistically (Agresti 1996):

- n_{AB}, the number of documents indexed with both term *A* and term *B*
- n_{Ab}, the number of documents indexed with term *A* but not term *B*
- n_{aB}, the number of documents indexed with term *B* but not term *A*
- n_{ab}, the number of documents indexed with neither term *A* or term *B*

In order to evaluate the statistical significance of the association between two MeSH terms, we use the likelihood ratio test (also called G-test or G-square test). The G^2 statistic compares the maximum of the likelihood function under two circumstances: 1) under the hypothesis of independence and 2) under the general, observed conditions. The G^2 statistic does not have the minimum expected frequency requirements imposed by the chi-square test. However, for the G^2 statistic to be computed, all observed frequencies must be greater than 0.

Because this study focuses on trans-ontological relations, statistically significant associations between MeSH terms are restricted to those pairs (*A*,*B*) where *A* and *B* belong to different MeSH hierarchies (e.g., between a disease in the C tree and an organism in the B tree).

Finally, in order to amplify the relations between disease categories and other categories in MeSH, the frequencies of co-occurrence are aggregated upwards, more precisely up to the second level of MeSH hierarchies. For example, the frequency of co-occurrence between *Cholera* [**C01**.252.400.959.347] and *Vibrio clolerae* [**B03**.440.450.900.859.225] is recorded as co-occurrence between *Bacterial Infections and Mycoses* [C01] and *Bacteria* [B03]. For those terms linked to more than one high-level category (e.g., *Lung Neoplasms*, linked to both *Neoplasms* [C04] and *Respiratory Tract Diseases* [C08]), the frequency of co-occurrence is divided among the corresponding high-level categories during the aggregation process, so as to avoid double-counting pairs.

5 Results

5.1 Expected Dependence Relations

Examples of dependence relations identified based on formal ontological principles among disease categories and other high-level categories in MeSH are presented below. Table 1 presents *has_participant* relations between pathological processes and anatomical entities, while Table 2 shows *has_agent* relations between pathological processes and pathogens.

Pathological process	**Anatomical entity**
Musculoskeletal Diseases	Musculoskeletal System
Digestive System Diseases	Digestive System
Stomatognathic Diseases	Stomatognathic System
Respiratory Tract Diseases	Respiratory System
Nervous System Diseases	Nervous System
Eye Diseases	Sense Organs (+)
Urological and Male Genital Diseases	Urogenital System
Female Genital Diseases and Pregnancy Complications	Urogenital System Embryonic Structures
Cardiovascular Diseases	Cardiovascular System
Hemic and Lymphatic Diseases	Hemic and Immune Systems
Skin Diseases	Integumentary System
Endocrine Diseases	Endocrine System

Table 1. Dependence relations of the type *has_participant* (between a pathological process and an anatomical entity).

Pathological Process	**Pathogen**
Bacterial Infection and Mycoses	Bacteria Fungi
Virus Diseases	Viruses
Parasitic Diseases	Animals (+)

Table 2. Dependence relations of the type *has_agent* (between a pathological process and a pathogen).

The MeSH descriptors marked with plus sign (+) represent cases where the corresponding term is too far down on the tree to be recognized here. For example, *Parasitic Diseases has_agent Animals* is true, but it would be more precise to say *has_agent Parasites*, *Parasites* being a subcategory of *Animals*.

5.2 Statistical Relations

25,376 pairs of co-occurring "starred" MeSH terms were extracted from the 385,491 citations in our 2004 MEDLINE set. The associations were statistically significant in all but 68 cases. Of these, 7,896 pairs of terms had a frequency of co-occurrence of at least 10 and were used in our analysis. In what follows, we report frequencies of co-occurrence after aggregation to the first subdivision of the MeSH trees, with emphasis on the associations between disease categories and other categories (6525 pairs).

Table 3 shows the number of associations between a given disease category and all other categories. Remarkably, for each disease category, there is generally one subcategory of the *Anatomy* and *Organisms* trees accounting for highest number of associations between this disease and other categories (e.g., between *Cardiovascular Diseases* and *Cardiovascular System*). Exceptions include *Neoplasms* [C04], *Congenital, Hereditary, and Neonatal Diseases and Abnormalities* [C16], *Endocrine Diseases* [C19] and *Immunologic Diseases* [C20].

Similarly, most *Anatomy* and *Organisms* categories are preferentially associated with one disease category. In contrast, other categories are frequently associated with most disease categories, but not one in particular, including *Pathological Conditions, Signs and Symptoms* [C23], *Amino Acids, Peptides, and Proteins* [D12], *Diagnosis* [E01], *Therapeutics* [E02] and *Surgical Procedures, Operative* [E04].

6 Discussion

6.1 Findings

Our hypothesis, i.e., that dependence relations are systematically corroborated by co-occurrence relations, has been for the most part verified in this experiment. In fact, among the 15 dependence relations identified between diseases and other categories, 11 (73%) correspond to the highest proportions of relations between this disease and any other category under investigation. The most remarkable exception is probably *Neoplasms* [C04]. The top category related to *Neoplasms* is *Amino Acids, Peptides, and Proteins* [D12], followed by *Surgical Procedures, Operative* [E04], which both correspond to contingent rather than dependence relations. As already mentioned, tumors are independent continuants, which helps to explain why there is not one particular category on which they depend. For *Endocrine Diseases* [C19], other dependence relations not studied here (namely to hormones) seem stronger than the dependence relation to *Endocrine System* [A06].

In some cases (e.g., *Otorhinolaryngologic Diseases* [C09]), the highest proportion of relations does not correspond to anatomical structures as expected, but to *Surgical procedures, Operative* [E04]. In fact, there is no such thing as an "otorhinolaryngologic system" and the corresponding anatomical structures (i.e., ear, nose and throat) are categorized in MeSH under *Respiratory System* [A04] and *Sense Organs* [A09]. These two categories, when taken together, correspond indeed to the highest proportion of relations for *Otorhinolaryngologic Diseases*. The same observation applies to groupings such as *Female Genital Diseases and Pregnancy Complications*, associated with *Urogenital System* and *Embryonic Structures*.

Of note, in most cases, each independent continuant such as anatomical entities and organisms is also preferentially associated with one disease category: the process which depends on this continuant. Exceptions include *Tissues* [A10], *Cells* [A11] and *Fluids and Secretions* [A12], all associated with several diseases. Finally, as expected, categories other than that predominantly associated with a disease generally stand in a contingent rather than dependence relation with this disease.

6.2 Significance

At first glance, the relations identified between diseases on the one hand and anatomical entities and organisms on the other appear essentially trivial. In fact, the curators of MeSH seem to purposely maintain a parallel in the names of these high-level descriptors (e.g., Virus Diseases / Viruses). However, none of the approaches to identifying relations presented in this study relies on the lexical properties of MeSH terms. Identifying dependence relations entirely relies on formal ontological principles and domain knowledge. The statistical relations were computed from MeSH identifiers rather than names. Moreover, relations between *Mycoses* and *Fungi* and between *Skin Diseases* and *Integumentary Sys-*

tem could probably not have been discovered by lexical analysis of the terms.

6.3 Limitations

For practical reasons, this statistical analysis is limited to co-occurrence data extracted from a subset of MEDLINE selected somewhat arbitrarily. The results would need to be confirmed on a larger subset.

Analogously, the statistical analysis of co-ocurrence data is the only technique used for identifying associations between MeSH terms in this study. Other techniques such as a vector space model and association rule mining could also be used and may yield different results (Bodenreider, Aubry and Burgun 2005). Our hypothesis should be tested against these different techniques as well.

Finally, we chose to amplify the results by aggregating frequencies of co-occurrence to the top-level categories in MeSH. While useful for this purpose, the aggregation prevented us from identifying fine associations. For example, as mentioned earlier, *Parasitic Diseases* [C03] is associated with *Animals* [B01], rather than *Parasites* [B01.500.714].

6.4 Applications

Semantic mining. The purpose of semantic mining is to identify and characterize the relations among entities of interest in a given domain. The statistical associations identified by most data mining techniques have to be characterized semantically in order to fully interpret the dataset. While this elucidation process is often part of the interpretation of the data, we argue that identifying dependence relations *a priori* may provide a useful framework for interpreting the relations discovered in a dataset.

Terminology creation and maintenance. Terminologies such as MeSH are built manually by experts and the relations explicitly represented in MeSH are limited to organizing descriptors in hierarchies. We showed that the approaches proposed in this paper can help identify and partially characterize trans-ontological relations. Representing dependence relations explicitly would undoubtedly facilitate terminology maintenance. Such relations could, for example, inform the MeSH editing environment that the modification of a descriptor must trigger the review of all descriptors standing in a dependence relation with the modified descriptor.

6.5 Generalizability

To other subdomains. This study was purposely limited to the relations between diseases and other categories. As mentioned earlier, many other categories are involved in dependence relations. Namely, all processes, not only pathological processes such as diseases, are dependent on some continuant. Physiology, for example, is a subdomain where many dependence relations can be identified.

To other terminologies. Similar to MeSH in many respects is the Gene Ontology™ (GO), a controlled vocabulary used to annotate (i.e., index) gene products across various model organisms. No trans-ontological relations are currently represented explicitly in GO. We showed in previous work that the approaches proposed in this paper would be applicable to GO as well (Bodenreider et al. 2005). While the statistical analysis of co-occurrence data requires a dataset, the formal ontological analysis can be applied to any terminology.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). This work was done while Lowell Vizenor was a visiting fellow at the Lister Hill National Center for Biomedical Communications, NLM, NIH. Our thanks go to Jim Mork and Kelly Zeng who helped extract the set of MEDLINE citations.

References

Agresti, A. 1996. *An introduction to categorical data analysis*. New York: Wiley.

Bodenreider, O., Aubry, M. and Burgun, A. 2005. Non-lexical approaches to identifying associative relations in the gene ontology. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung and T. E. Klein (eds.). *Pacific symposium on biocomputing 2005*, pp. 91-102. World Scientific.

Grenon, P., Smith, B. and Goldberg, L. 2004. Biodynamic ontology: Applying BFO in the biomedical domain. In D. M. Pisanelli (ed.). *Ontologies in medicine*, pp. 20-38. Amsterdam: IOS Press.

Rosse, C., Kumar, A., Mejino, J. L. V., Cook, D. L., Detwiler, L. T. and Smith, B. 2005. A strategy for improving and integrating biomedical ontologies. *Proc AMIA Symp*: 639-643.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L. and Rosse, C. 2005. Relations in biomedical ontologies. *Genome Biol* 6(5): R46.

Smith, B. and Grenon, P. 2004. The cornucopia of formal ontological relations. *Dialectica* 58(3): 279-296.

		C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	F03	
		Bacterial Infections and Mycoses	Virus Diseases	Parasitic Diseases	Neoplasms	Musculoskeletal Diseases	Digestive System Diseases	Stomatognathic Diseases	Respiratory Tract Diseases	Otorhinolaryngologic Diseases	Nervous System Diseases	Eye Diseases	Urologic and Male Genital Diseases	Female Genital Diseases and Pregnancy Complications	Cardiovascular Diseases	Hemic and Lymphatic Diseases	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Skin and Connective Tissue Diseases	Nutritional and Metabolic Diseases	Endocrine Diseases	Immunologic Diseases	Disorders of Environmental Origin	Animal Diseases	Mental Disorders	*TOTAL*
A01	Body Regions	0.00	0.00	0.00	6.75	1.00	3.50	0.00	0.00	0.00	1.50	1.50	0.50	0.00	12.00	0.00	1.00	1.00	2.00	3.25	0.00	1.00	0.00	0.00	35.00
A02	Musculoskeletal System	1.50	0.00	0.00	6.17	52.08	0.00	2.58	1.75	0.00	11.17	1.33	0.50	0.17	5.33	0.00	1.50	1.42	0.75	0.92	1.75	36.75	0.50	0.00	126.17
A03	Digestive System	1.50	1.50	0.00	18.83	0.00	45.75	0.33	0.17	0.00	0.17	0.00	0.50	1.00	5.00	0.00	0.00	0.00	4.69	1.11	0.11	5.83	0.00	0.00	86.50
A04	Respiratory System	0.50	0.50	0.00	3.58	0.25	1.33	0.00	21.25	4.00	0.50	0.00	0.00	0.00	2.50	0.00	0.83	0.00	0.00	0.00	2.08	2.33	1.00	0.00	40.67
A05	Urogenital System	0.00	1.00	0.00	9.08	0.00	1.00	0.00	0.00	0.00	0.50	0.00	31.92	9.67	3.50	0.00	2.17	0.50	0.50	2.00	0.00	1.50	0.00	0.50	63.83
A06	Endocrine System	0.00	0.00	0.00	3.50	0.00	0.00	0.00	0.00	0.00	0.67	0.33	1.33	1.17	0.00	0.00	0.17	0.00	0.94	7.11	0.94	0.00	0.00	0.00	16.17
A07	Cardiovascular System	1.00	0.00	0.00	5.17	0.00	2.50	0.00	0.50	0.00	4.92	1.00	1.17	1.42	101.00	0.00	4.75	0.00	2.00	0.75	0.00	0.67	0.00	0.00	126.84
A08	Nervous System	0.00	0.75	0.00	4.67	0.00	0.00	0.50	0.50	0.00	84.33	4.92	0.00	0.00	13.50	0.00	5.25	0.00	1.00	0.17	2.17	5.75	1.50	26.50	151.50
A09	Sense Organs	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	3.00	3.42	34.58	0.00	0.00	1.67	0.00	0.00	0.00	0.50	1.17	0.00	2.00	0.00	0.00	46.83
A10	Tissues	1.25	0.00	0.00	13.00	7.08	6.75	2.08	1.83	0.58	4.25	4.50	1.25	1.67	10.50	1.00	0.33	1.75	5.50	3.00	1.00	6.42	0.75	0.00	74.50
A11	Cells	0.50	1.00	0.75	10.67	5.00	2.50	0.00	1.00	0.00	10.17	2.17	1.50	0.00	11.50	1.67	0.50	4.00	4.28	1.44	7.53	3.08	0.00	1.75	71.00
A12	Fluids and Secretions	3.50	0.00	0.00	1.33	0.17	1.83	2.50	2.83	0.00	0.50	1.00	1.83	1.33	0.00	0.00	0.83	0.17	0.00	0.25	0.67	0.00	1.75	0.00	20.50
A13	Animal Structures	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.25	0.00	0.00	3.25	0.00	4.50
A14	Stomatognathic System	0.00	0.00	0.00	0.75	1.92	0.00	15.50	1.17	0.00	1.17	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	2.83	0.00	0.00	23.83
A15	Hemic and Immune Systems	1.00	1.00	0.75	11.08	0.17	0.50	0.25	1.25	0.00	0.50	0.00	0.00	0.50	2.00	2.58	0.00	1.92	0.33	1.08	6.58	1.00	0.00	0.00	32.50
A16	Embryonic Structures	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	5.25	0.00	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.50
A17	Integumentary System	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.33	4.83	0.00	0.00	0.33	1.00	0.00	0.00	9.00
B01	Animals	2.83	4.12	54.50	0.53	0.00	3.17	0.00	0.00	0.00	0.00	0.00	0.33	0.50	0.00	0.70	0.00	4.00	0.00	0.00	0.28	6.00	32.20	0.00	109.17
B02	Algae	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.17
B03	Bacteria	109.00	0.08	0.00	1.83	0.50	9.33	3.33	5.17	1.33	1.00	0.00	1.00	0.50	1.00	1.33	0.33	2.00	0.00	0.50	0.42	0.50	11.00	0.00	150.17
B04	Viruses	2.50	66.00	0.50	13.33	0.00	9.50	0.33	5.00	0.33	2.75	0.00	1.67	2.17	1.00	1.17	0.00	2.00	1.00	0.00	8.08	0.50	11.58	0.75	130.17
B05	Fungi	12.50	0.08	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.58	1.50	0.00	0.00	18.17
B06	Plants	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.33	0.00	0.00	0.00	3.00
B07	Archaea	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.17
C23	Pathological Conditions, Signs and Symp	19.33	2.50	0.00	68.08	6.17	33.58	0.25	11.50	1.42	27.98	12.08	9.58	5.78	52.29	3.56	4.46	3.75	7.75	7.07	4.35	7.83	3.75	9.33	302.38
D01	Inorganic Chemicals	0.00	0.00	0.00	13.63	2.00	1.00	2.00	3.25	0.00	1.50	0.33	2.83	1.17	6.63	2.46	0.50	0.50	5.00	4.67	1.54	5.50	0.00	2.50	57.00
D02	Organic Chemicals	6.75	2.25	6.00	21.83	9.50	3.75	0.00	5.75	0.00	12.50	2.50	5.33	2.50	22.00	0.58	0.33	1.92	7.25	7.50	7.00	6.75	0.00	17.25	149.25
D03	Heterocyclic Compounds	13.00	7.33	4.00	35.63	1.17	13.75	0.00	4.17	0.17	9.08	3.00	4.08	2.17	29.13	2.71	0.75	1.92	7.75	3.67	7.04	7.00	0.00	24.25	181.75
D04	Polycyclic Hydrocarbons	0.00	0.50	0.00	3.17	0.00	1.50	1.00	3.17	0.17	0.75	0.83	1.75	0.17	5.50	0.17	0.00	0.33	3.00	3.25	1.92	1.00	0.00	1.50	29.67
D06	Hormones, Hormone Substitutes, and Hc	0.67	0.00	0.00	4.00	1.78	0.00	0.00	2.08	0.00	1.03	0.17	0.92	0.50	3.00	0.17	0.00	0.00	5.83	8.44	1.42	0.00	0.00	0.50	30.50
D07	Reproductive Control Agents	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50
D08	Enzymes and Coenzymes	5.33	1.50	0.00	32.86	0.33	5.17	0.00	3.67	0.00	4.92	0.00	2.92	1.28	6.17	1.42	1.17	0.58	2.00	3.53	2.33	4.00	0.00	1.00	80.17
D09	Carbohydrates and Hypoglycemic Agents	4.75	0.25	1.00	7.58	2.17	0.50	0.00	1.75	0.00	1.00	0.50	0.25	0.50	11.25	1.00	0.00	0.17	17.08	7.67	1.08	0.00	0.00	0.00	58.50
D10	Lipids	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.00	0.00	1.50	1.50	0.00	0.00	11.50	0.00	0.00	0.33	13.75	1.58	0.00	0.00	0.00	0.50	31.33
D11	Growth Substances, Pigments, and Vitan	0.00	0.00	0.00	5.56	0.00	1.13	0.00	0.54	0.00	0.25	1.17	0.67	0.17	1.23	0.06	0.75	0.83	2.50	1.08	0.06	0.00	0.00	0.00	16.00
D12	Amino Acids, Peptides, and Proteins	10.50	8.75	1.00	119.78	6.89	28.25	1.83	8.75	1.83	30.14	3.00	11.50	7.28	46.83	11.42	11.75	8.42	23.17	25.83	17.58	7.08	2.00	12.25	405.83
D13	Nucleic Acids, Nucleotides, and Nucleosi	0.50	9.75	0.00	16.33	0.00	6.50	0.33	1.00	0.33	0.25	0.25	0.83	1.67	0.00	0.33	0.25	0.00	1.25	1.33	5.08	3.00	0.00	0.00	49.00
D14	Neurotransmitters and Neurotransmitter /	0.67	0.00	0.00	1.58	0.22	0.00	0.00	2.17	0.17	3.89	0.00	1.33	0.00	7.58	0.00	0.00	0.00	0.17	0.72	1.17	0.75	0.00	5.67	26.08
D15	Central Nervous System Agents	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.33	0.00	14.33	0.25	0.00	0.50	1.50	0.33	0.50	0.00	2.00	0.50	0.00	4.50	0.00	15.75	41.33
D16	Peripheral Nervous System Agents	0.33	0.00	0.00	0.50	1.03	1.83	0.00	0.67	0.00	2.42	0.00	0.33	0.00	0.33	0.33	0.00	0.72	0.00	0.00	0.89	0.36	0.00	0.25	10.00
D17	Anti-Inflammatory Agents, Antirheumatic	0.33	0.00	0.00	0.17	1.78	1.83	0.00	0.92	0.00	0.33	0.50	0.33	0.00	1.67	0.00	0.00	1.22	0.00	0.00	1.47	0.11	0.00	0.25	10.92
D18	Cardiovascular Agents	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.42	1.50	1.75	0.00	18.50	0.00	0.00	0.00	0.50	0.75	0.00	0.00	0.00	0.25	24.33
D19	Hematologic, Gastrointestinal, and Renal	0.75	0.00	0.00	0.50	0.00	3.00	0.00	0.42	0.00	0.67	0.00	0.25	0.25	12.50	2.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	20.83
D20	Complex Mixtures	0.00	0.00	0.00	0.58	0.00	0.83	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.42	0.00	1.17	0.00	0.00	3.50
D21	Anti-Allergic and Respiratory System Age	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.42	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	1.00	0.00	0.00	0.00	4.00
D22	Antineoplastic and Immunosuppressive A	0.00	0.00	0.00	49.65	0.67	8.63	1.25	1.63	0.00	1.25	0.00	3.42	2.67	1.31	5.15	0.17	2.25	0.00	2.58	4.40	0.50	0.00	0.00	85.50
D23	Dermatologic Agents	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00	3.50
D24	Immunologic and Biological Factors	17.17	29.75	6.50	67.59	3.98	21.13	0.70	11.46	0.33	7.19	0.53	9.67	6.28	12.06	7.81	2.50	12.62	4.92	7.64	28.85	4.00	3.00	1.00	266.67

Table 3. Frequency of co-occurrence between disease categories and other high-level categories in MeSH (1/2)

		C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	F03	
		Bacterial Infections and Mycoses	Virus Diseases	Parasitic Diseases	Neoplasms	Musculoskeletal Diseases	Digestive System Diseases	Stomatognathic Diseases	Respiratory Tract Diseases	Otorhinolaryngologic Diseases	Nervous System Diseases	Eye Diseases	Urologic and Male Genital Diseases	Female Genital Diseases and Pregnancy Complications	Cardiovascular Diseases	Hemic and Lymphatic Diseases	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Skin and Connective Tissue Diseases	Nutritional and Metabolic Diseases	Endocrine Diseases	Immunologic Diseases	Disorders of Environmental Origin	Animal Diseases	Mental Disorders	*TOTAL*
D25	Biomedical and Dental Materials	0.00	0.25	0.00	0.00	1.50	0.25	4.33	0.00	0.00	0.00	0.50	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.50	0.00	0.00	10.33
D26	Pharmaceutical Preparations	0.00	0.00	0.00	2.00	0.00	0.50	0.00	0.00	0.00	0.33	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.00	0.67	0.00	0.50	4.83
D27	Chemical Actions and Uses	47.42	14.50	10.33	61.27	7.78	21.29	4.58	17.04	3.17	22.25	8.25	14.08	6.58	67.35	8.85	1.25	10.47	15.08	12.50	16.58	21.11	2.50	21.42	415.67
E01	Diagnosis	4.30	2.03	0.00	63.54	6.42	21.96	0.50	12.46	1.78	36.72	10.53	7.27	14.16	71.48	0.25	9.74	2.42	3.60	8.85	5.73	4.10	1.00	6.75	295.58
E02	Therapeutics	13.00	7.75	0.25	66.26	3.92	17.63	0.00	17.71	0.78	24.00	3.00	18.08	8.50	62.15	5.29	5.03	3.08	8.86	12.22	11.32	6.50	0.75	2.50	298.58
E03	Anesthesia and Analgesia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.50	0.00	0.00	1.00	2.50	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	10.50
E04	Surgical Procedures, Operative	8.00	6.25	0.00	79.06	22.50	40.08	4.08	10.08	6.86	14.83	28.00	17.83	14.00	77.28	3.42	9.28	5.92	6.08	12.78	7.25	38.50	0.00	0.50	412.58
E05	Investigative Techniques	4.22	3.86	0.00	32.03	2.33	5.84	0.58	5.05	0.64	5.13	0.36	3.78	3.16	12.59	0.25	0.77	0.58	2.77	3.66	3.35	1.90	3.25	2.83	98.94
E06	Dentistry	0.00	0.00	0.00	0.00	2.92	0.00	22.25	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.25	0.00	0.00	26.08
E07	Equipment and Supplies	9.25	1.50	0.50	2.92	6.75	8.00	3.83	1.50	2.50	3.25	5.50	4.67	1.17	42.50	0.00	0.50	2.25	0.17	0.42	2.67	19.00	0.00	0.00	118.84
F01	Behavior and Behavior Mechanisms	1.42	7.96	0.00	14.33	0.50	0.75	1.50	3.42	0.00	15.10	0.00	0.50	3.00	12.92	0.00	0.25	0.00	10.11	5.11	7.57	11.33	0.00	54.60	150.37
F02	Psychological Phenomena and Processe	0.00	0.17	0.00	2.13	1.50	0.00	0.00	0.25	1.00	22.33	1.00	0.50	0.00	3.75	0.13	0.13	0.00	1.00	0.88	0.17	2.17	0.25	21.54	58.87
F04	Behavioral Disciplines and Activities	0.00	0.42	0.00	1.00	0.00	0.00	0.00	0.00	0.00	5.83	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.42	0.92	0.00	40.00	48.92
G01	Biological Sciences	0.13	0.00	0.00	1.98	0.00	0.43	0.00	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.24	0.00	0.00	0.79	0.00	0.38	0.00	0.00	4.03
G02	Health Occupations	0.92	0.25	0.08	4.83	0.83	1.00	0.50	0.67	1.00	2.00	1.00	2.00	3.00	3.00	1.00	0.00	2.33	1.83	1.33	1.33	3.00	1.08	3.25	36.25
G03	Environment and Public Health	31.01	19.90	6.83	12.81	0.11	5.92	0.92	9.96	1.70	3.38	0.25	0.62	1.25	4.95	0.00	0.41	1.86	3.68	1.52	11.50	32.19	6.08	7.50	164.36
G04	Biological Phenomena, Cell Phenomena,	14.33	2.71	0.33	25.00	0.75	3.17	0.75	1.00	0.00	2.92	0.17	2.00	0.00	3.75	1.92	0.00	4.00	0.33	1.75	5.63	6.42	0.33	0.00	77.25
G05	Genetic Processes	0.00	0.00	0.00	26.08	0.00	4.33	0.50	1.17	0.00	1.17	0.00	1.83	1.94	0.75	0.47	0.67	0.00	0.00	2.28	0.47	4.00	0.00	0.17	45.83
G06	Biochemical Phenomena, Metabolism, ar	0.00	0.75	0.00	7.44	0.00	1.92	0.00	0.33	0.00	2.25	0.00	1.42	0.11	5.83	0.00	0.33	0.00	8.75	2.53	0.75	1.50	0.00	1.25	35.17
G07	Physiological Processes	0.00	0.50	0.00	2.36	0.17	0.50	0.17	0.58	0.00	3.67	1.00	0.75	0.44	3.50	0.00	0.17	0.00	6.42	1.86	0.67	1.50	1.00	3.83	29.08
G08	Reproductive and Urinary Physiology	0.67	1.83	0.00	1.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	3.83	9.00	1.25	0.25	2.50	0.00	1.00	1.50	1.42	0.25	0.50	2.00	29.00
G09	Circulatory and Respiratory Physiology	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.50	0.33	3.33	0.00	1.00	0.00	40.83	0.00	0.00	0.00	0.50	0.50	0.29	0.50	0.00	0.00	47.92
G11	Musculoskeletal, Neural, and Ocular Phys	0.00	0.00	0.00	0.63	7.00	0.00	0.42	2.58	1.00	17.99	9.83	0.00	0.00	7.33	0.13	0.29	0.00	1.94	0.57	0.11	4.67	0.25	3.13	57.87
G12	Chemical and Pharmacologic Phenomen	8.00	0.50	0.00	5.17	0.00	1.50	0.00	0.75	0.00	0.00	0.00	0.75	0.17	1.00	0.00	0.00	0.00	1.75	1.17	0.50	0.00	0.00	0.00	21.25
G13	Genetic Phenomena	0.00	1.00	0.00	23.10	2.50	8.42	0.00	2.58	1.50	8.42	3.00	1.08	2.17	6.13	2.01	6.50	2.00	3.25	5.58	4.26	1.50	0.83	6.17	92.00
G14	Genetic Structures	0.00	0.00	0.00	18.50	0.00	3.25	0.00	0.50	0.00	0.25	0.00	1.33	1.00	0.25	0.00	0.75	0.00	0.00	2.92	0.00	0.33	0.00	0.25	29.33
H01	Natural Sciences	0.13	0.00	0.00	23.14	1.50	3.51	0.50	2.71	1.00	4.42	0.78	1.60	0.83	6.29	0.00	0.24	0.50	0.50	2.36	0.00	6.93	0.50	0.75	58.17
I01	Social Sciences	2.67	7.29	0.00	6.33	0.28	2.75	0.00	1.58	0.00	5.58	0.00	1.25	1.00	1.75	0.25	0.25	0.78	2.94	1.36	7.18	8.58	0.00	28.58	80.42
I02	Education	0.00	0.58	0.00	1.33	0.00	0.00	0.33	0.75	0.00	0.50	0.00	0.25	0.00	1.83	0.00	0.00	0.00	3.00	0.92	1.00	0.42	0.00	1.25	12.17
I03	Human Activities	1.00	0.50	1.00	0.00	0.33	0.00	0.00	0.83	0.00	0.67	0.00	0.00	0.00	3.50	0.00	0.00	0.00	2.44	0.44	0.11	4.50	0.00	2.50	17.83
J01	Technology, Industry, and Agriculture	0.13	0.00	0.50	0.50	0.50	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0.75	0.50	8.38	0.00	0.00	13.00
J02	Food and Beverages	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.50	0.00	1.00	0.00	0.00	0.00	2.75	1.00	0.00	0.00	3.25	0.00	9.50
K01	Humanities	0.00	0.00	0.00	3.00	0.17	2.75	0.00	1.25	0.00	4.00	0.00	1.25	0.00	0.75	0.25	0.25	0.67	0.25	0.50	0.67	0.25	0.00	2.50	18.50
L01	Information Science	0.00	0.00	0.00	7.72	0.00	0.00	0.00	0.63	0.00	5.25	0.00	0.50	0.50	3.48	0.00	0.00	0.00	1.50	1.85	0.13	2.00	0.33	5.08	28.97
M01	Persons	1.50	6.25	0.00	6.83	0.00	1.00	0.00	1.67	0.50	4.25	1.50	2.50	2.50	4.75	0.00	3.00	0.00	4.61	2.61	4.53	7.42	0.00	18.42	73.83
N01	Population Characteristics	1.92	4.42	0.08	1.75	1.11	0.17	1.00	1.50	0.00	0.83	0.00	0.00	0.83	1.75	0.00	0.17	0.11	2.67	0.50	2.78	3.00	0.08	4.00	28.67
N02	Health Care Facilities, Manpower, and Se	4.30	7.65	0.00	11.46	0.00	2.84	0.83	4.59	0.20	9.97	0.00	0.85	2.00	10.70	0.00	1.91	0.00	4.27	3.49	4.93	11.18	0.75	16.42	98.33
N03	Health Care Economics and Organizatior	0.00	4.04	0.00	1.50	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.50	0.67	4.04	1.42	0.00	2.17	15.59
N04	Health Services Administration	0.25	1.50	0.00	4.85	0.00	0.00	0.00	2.50	0.00	2.75	0.00	0.10	0.00	8.75	0.00	0.00	1.00	4.25	4.20	2.75	4.22	0.00	10.17	47.28
N05	Health Care Quality, Access, and Evaluat	2.22	5.86	0.00	11.23	0.50	1.21	0.58	2.26	0.20	4.88	0.25	0.37	1.00	8.87	0.00	0.41	0.00	2.93	4.19	5.35	1.65	0.00	11.33	65.27
	TOTAL	374.9	249.1	95.4	1097.5	178.2	381.3	83.4	218.8	38.5	498.2	156.9	193.7	139.1	907.3	72.4	87.8	101.1	240.0	216.9	238.6	356.8	95.0	403.1	6325.00

Table 3. Frequency of co-occurrence between disease categories and other high-level categories in MeSH (2/2)