Therapeutic Category Improvement Method Based on the Words Appearing in Effect-Efficacy Description

Hirotsugu Ishida[1], Keita Nabeta[1], Masaomi Kimura[2], Michiko Ohkura[2], and Fumito Tsuchiya[3]

[1] Graduate School of Engineering, Shibaura Institute of Technology, 3-7-5 Toyosu, Koto Ward, Tokyo 135-8548 Japan
[2] Shibaura Institute of Technology, 3-7-5 Toyosu, Koto Ward, Tokyo 135-8548 Japan
[3] International University of Health and Welfare, 2600-1 Kitakanemaru, Ohtawara City, Tochigi, Japan
{m110013,m709102}@shibaura-it.ac.jp, {masaomi,ohkura}@sic.shibaura-it.ac.jp, ftsuchiya@iuhw.ac.jp

Abstract. Medical drugs have various efficacies, and are classified focusing on their purpose of use. In Japan, the Ministry of Internal Affairs and Communications gives Japan standard commodity classification (JSCC) numbers to drugs. Therapeutic category numbers are decided based on three digit numbers after the head digits "87". Although the current JSCC numbers are determined based on the revised document "Japan standard commodity classification" compiled in 1990, they have not been revised for 20 years. As a result, when drugs are categorized based on this categorizing system, some drugs are not applicable to any category. As the result, the drugs have been categorized as "other categories" such as "drug for other allergy" or "drug for other cardiovascular disease." The number of such drugs is increasing. However, since it is conceivable that drugs having similar efficacy are often included in other categories, it is necessary that such drugs are classified independently from the "other categories." Therefore, in this study, we analyzed drugs information categorized as "drugs for other cardiovascular disease," and proposed a method of classifying these drugs by using clustering.

Keywords: Medical Safety, Therapeutic Category, Clustering.

1 Introduction

Medical drugs have various efficacies, and they are classified focusing on the purpose of use. In Japan, the Ministry of Internal Affairs and Communications [1] gives Japan standard commodity classification (JSCC) numbers to drugs, which are composed of 5 or 6 digits. Ethical and proprietary drugs are given JSCC numbers whose head two digits are "87," then therapeutic category numbers are decided based on three digits of numbers after "87." Therapeutic category numbers have a hierarchical structure [2], and the third digit following the two digits "87" expresses the action part of the body

J.A. Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2011, LNCS 6764, pp. 174–181, 2011.

or the efficacy. The fourth digit expresses an ingredient or the action part of the body, and the fifth digit expresses the use. For example in the case of "87214," the third digit "2" expresses "drug for individual organic systems," the fourth digit "1" expresses "drug for cardiovascular disease," and the fifth digit "4" expresses "a hypotensive drug."

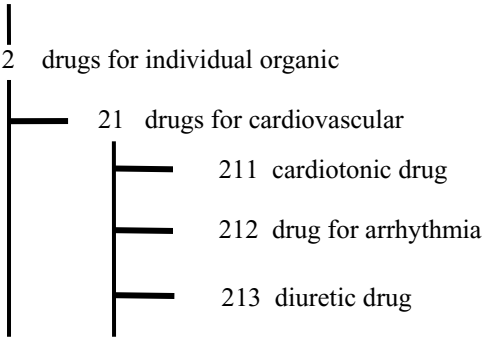

Fig. 1. Part of the hierarchical structure of the therapeutic category numbers

Although the current JSCC numbers are determined based on the revised document "Japan standard commodity classification" compiled by the Ministry of Internal Affairs and Communications, they have not been revised for 20 years. As a result, when drugs are categorized based on this categorizing method, some drugs are not applicable to any category. Therefore, the drugs have been categorized as the "other category" such as "drug for other allergy" or "drug for other cardiovascular disease." The number of such drugs is increasing.

However, since it is conceivable that drugs having similar efficacy are often included in other categories, it is necessary that such drugs are classified independently from the "other categories." Therefore, in this study, we analyzed drugs information categorized as "drugs for other cardiovascular disease," and proposed a method of classifying these drugs by using clustering.

2 Target Data

We obtained 166 drug package inserts formatted with Standard General Markup Language (SGML) whose therapeutic category number is "219," which corresponds to "drug for other cardiovascular disease," from the Pharmaceuticals and Medical Devices Agency (PMDA) [3]. Package inserts are exclusive legal documents used to describe detailed information for each drug, including the composition, efficacy, dosage and cautions. As plural drugs are described in one package insert for every standard, we targeted 235 drugs.

3 The Method Based on Therapeutic Category Names

Firstly, we obtain the therapeutic category names from each drug. Therapeutic category names are defined as "if you can express definitely the efficacy or the character of such drug, you mention them and avoid expressions that may invite misunderstanding by the user" in the notice [4] given in 1997. As such therapeutic category names are decided by each pharmaceutical company, there is no standardization. As a result, there are therapeutic category names such as "circulatory disease improvement drug" based on the concrete efficacy of the drugs and "prostaglandin E1 drug" based on ingredient names of the drugs. Although we can find the efficacy of the drugs directly based on the concrete efficacy as in the former case, we cannot find the efficacy of the drugs directly based on only ingredient names as in the latter case. In addition, some drugs are not given therapeutic category names. Therefore, we cannot classify drugs having similar efficacy based on therapeutic category names.

4 Proposed Method

As mentioned above, since we cannot classify drugs based on therapeutic category names, we classify drugs based on information described on the part of "effect-efficacy" in the package inserts. We generate networks based on the words included in the part of "effect-efficacy" and aggregate drugs with similar efficacy by applying a clustering method to the networks.

4.1 Extraction of Nouns

We divide the statements of "effect-efficacy" into clauses with "CaboCha," which is the Japanese dependency analyzer, and extract nouns included in these clauses. In addition, if we extract compound words that plural nouns connect like "頭部外傷後遺症 (head injury aftereffects)," we divide them according to their meanings like "頭部(head)," "外傷(injury)," "後遺症(aftereffect)" and extract these nouns.

Then we count up the number of nouns that are extracted by the method for every drug. Nouns such as "改善(improvement)," "下記(follows)," "もの(thing)" and so forth appear for many drugs. Since we expect that they are not related with the effect directly, we exclude such nouns from the target of analysis. In addition, we exclude nouns that are provided by dividing the compound words and not related with the effect directly. For example, we exclude "慢性(chronic)," which was provided by dividing "慢性腎不全(chronic renal insufficiency)" and leave "腎不全(renal insufficiency)."

4.2 Generation of Networks

We make a connection matrix with the remaining nouns and the product names as a network with them as Fig. 2. A row of the connection matrix expresses what kind of

nouns appear in statements of "effect-efficacy" of the package insert of a drug. If a noun appears in a drug statement, the element falling under the noun of the row of the drug is 1. If the noun does not appear in the drug statement, it is 0. Then we make an adjacency matrix between the product names based on the connection matrix as Fig. 3. If drugs connect through common nouns in the connection matrix, we set up an edge between the drugs in the adjacency matrix. Therefore, the adjacency matrix expresses the network of the product name of drugs.

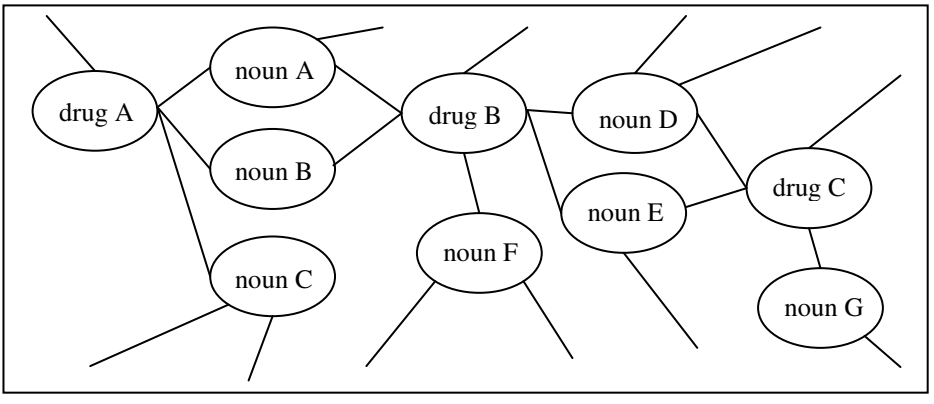

Fig. 2. An example of a network with drugs and nouns

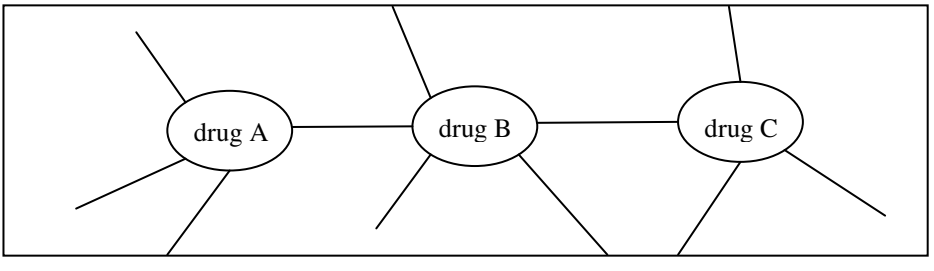

Fig. 3. An example of a network of drugs

5 Analysis 1

5.1 Method

We apply the non-hierarchical clustering method by the modularity [5] to the adjacency matrix and extract clusters connected as the same efficacy. The modularity is one of the indexes for detecting the communities from the network. If a connection

between nodes in a community in the network is dense and a connection between the communities is sparse, the modularity takes the high value. This method using modularity classifies the drugs based on the network of the drugs.

5.2 Results

As a result of non-hierarchical clustering, the drugs were divided into three clusters. The numbers of drugs that belong to each cluster were 104, 72 and 57. Table 1 shows the top five frequencies of nouns that appear in the drugs included in each cluster.

Table 1. Frequencies of appearance of nouns in the top five cases in each cluster

	Cluster 1		Cluster 2		Cluster 3	
	Noun	Frequency	Noun	Frequency	Noun	Frequency
1	脳 (brain)	99	血症 (blood symptom)	45	循環 (circulation)	48
2	梗塞 (infarction)	91	血圧 (blood pressure)	43	閉塞性 (obstructive)	48
3	障害 (failure)	70	症 (symptom)	28	動脈 (artery)	48
4	頭部 (head)	55	透析 (dialysis)	28	硬化症 (sclerosis)	48
5	外傷 (injury)	55	腎不全 (kidney failure)	28	障害 (failure)	42

5.3 Discussion

Since many nouns about the head, such as “brain” “head” and so forth, appeared in Cluster 1, which is the largest cluster, we expected that Cluster 1 would be independent as circulatory organs drugs related to the head.

However, because the number of drugs in Cluster 2 is 72, and the frequency of the nouns that appear most in the cluster is 45, it is desirable to subdivide Cluster 2 including the small clusters that we should subdivide in Cluster 2. Therefore, since we cannot subdivide Cluster 2 by using the non-hierarchical clustering method, we apply the hierarchical clustering method for analysis based on inclusion relations between the clusters.

6 Analysis 2

6.1 Method

We apply the hierarchical clustering method for the connection matrix with the remaining nouns and the product names made in Section 4.2. We use the Euclid distance for distances between the drugs. Then, in order to demand a cluster expecting that it has drugs having similar efficacy, we calculate the entropy using the following expressions after the cluster division. The entropy is an index to express disorder of the information.

$$S = \sum_{i=1}^{M} \sum_{W} P_{con}(w) \times (R_W^i \log R_W^i + (1 - R_W^i) \log(1 - R_W^i)) \quad (1) \tag{1}$$

R_W^i is a ratio of the elements including noun "w" in cluster i[th] after the division, and $P_{con}(w)$ is a ratio of the number of noun "w" for that of all nouns. "M" is the number of clusters. If the number of the clusters is two after the division, we calculate the entropy as M=2. If the number of the clusters is 1 before the division, we calculate the entropy as M=1. Then we calculate the information gain. The larger the information gain, the better the division. Therefore, we repeatedly divide the clusters until the information gain decreases in comparison with the division before 1.

6.2 Results

We show the results in Fig. 4 and Table 2. Figure 4 is a dendrogram showing the information gains of each division. Table 2 shows the number of elements and therapeutic category names of drugs in each cluster. As a result of the hierarchical clustering, the drugs were divided into five clusters. The numbers of drugs belonging to each cluster were 154, 32, 32, 8 and 8.

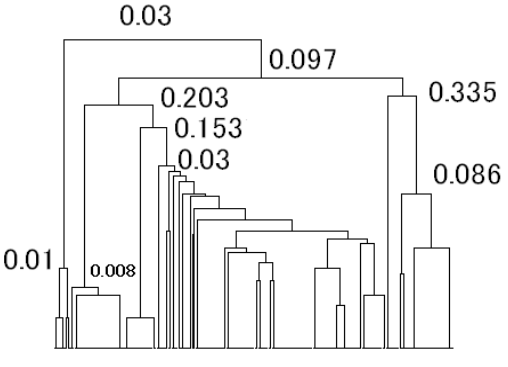

Fig. 4. Results of the hierarchical clustering method. The numbers in the figure show information gains.

Table 2. The number of elements and therapeutic category names of drugs in each cluster

Cluster number	Number of elements	Therapeutic category names of drugs in each cluster
1	154	<include many drugs with different efficacy>
2	32	"disturbance of consciousness / pancreatitis treatment drug" "Brain metabolism improvement drug" etc.
3	32	"prostaglandin E1 drug" etc.
4	9	"antithrombin drug" etc.
5	8	"circulatory disease improvement drug" "antihypertensive drug" etc.

6.3 Discussion

Since drugs having various different efficacies still coexisted in Cluster 1 including 154 elements, we could not extract therapeutic category names that plainly express the efficacy of drugs in the cluster. On the other hand, since the other clusters have therapeutic category names having almost the same meaning in each cluster, we could extract the efficacy of drugs in each cluster. According to a result such as "circulatory disease improvement drug" based on concrete efficacy or "prostaglandin E1 drug" based on the ingredient name of the drug, therapeutic category names are decided from plural viewpoints, as a therapeutic category itself. Therefore, it is necessary to unify these viewpoints of therapeutic category.

7 Conclusion

In this study, we analyzed the classification system focusing on "drugs for other cardiovascular disease" to contribute to the improvement of therapeutic categories, which have not been revised for 20 years. In order to classify drugs having similar efficacy, we proposed a method of clustering the connection of the product names of drugs and nouns included in the statements of "effect-efficacy." When we applied a non-hierarchical clustering method, there was a cluster that should be subdivided because the cluster includes small clusters. However, since we could not subdivide such cluster by the non-hierarchical clustering method, we applied a hierarchical clustering method that performs analysis based on inclusion relations between the clusters.

As a result, we classified drugs into clusters including drugs of similar efficacy. However, therapeutic category names are decided from plural viewpoints, as a therapeutic category itself. Therefore, it is necessary to unify these viewpoints of therapeutic category. In future, it is necessary to analyze the other categories by applying our method and propose a new classification system.

References

1. Ministry of Internal Affairs and Communications (2011), http://www.stat.go.jp/index/seido/syouhin/gaiyou.htm
2. Mochidsuki, M.: Revision; How to read package inserts to understand drugs definitely, Jiho (2004)
3. Pharmaceuticals and Medical Devices Agency (2011), http://www.info.pmda.go.jp/
4. Notification of the Director of the Ministry of Health and Welfare Pharmaceutical Practice: The meanings of statements in ethical drug package inserts (1997)
5. Reichardt, J., Bornhold, S.: Statistical mechanics of community detection. Physical Review E 74,016110, 1–14 (2006)
6. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69, 026113 (2004)
7. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E 69, 066133 (2004)
8. Aaron Clauset, M.E.J.: Newman and Christopher Moore: Finding community structure in very large networks. Physical Review E 70, 066111 (2004)