Fractals, Vol. 10, No. 3 (2002) 341–351

CLINICAL DIAGNOSES FOLLOWING ZIPF'S LAW

YUTAKA TACHIMORI
Department of Anesthesiology
Hanwa dai-2 Senboku Hospital, 3176 Hukai
Kitamachi, Sakai, Osaka, 559-8271, Japan
tachimori@tcct.zaq.ne.jp

TAKASHI TAHARA
Department of Medical Economy
Nihon Fukushi University, Okuda
Mihama-cho, Chita-gun, Aichi, 470-3295, Japan

Received June 11, 2001; Accepted September 5, 2001

Abstract

We examine the statistical properties of data relating to clinical diagnoses extracted from the medical database of our hospital. Specifically, we analyze all diagnoses given to all patients (in- and out-patients). The data consisted of both data input to computers as Japanese names, and the data input as ICD10[a] codes (International Classification of Diseases 10th Revision). We adapted the Zipf approach to analyzing the frequencies of clinical diagnoses for these two data groups. We found that both group types have the inverse-power relationship between the rank order of diagnoses and the frequency of the appearance of these diagnoses. (This relationship is called Zipf's law, which is observed in natural language.) Though the reason why these sets follow Zipf's law is unknown, we speculate that the complex interaction between doctor and patient is the cause for adherence to Zipf's law.

Keywords: Zipf's Law; Complex System; Clinical Diagnosis; Medical Structure; Fractal.

[a]The International Classification of Diseases is a system of categories for classifying various forms of morbidity. The ICD is designed to facilitate the statistical study of disease phenomena.

1. INTRODUCTION

In medical treatment, clinical diagnosis can be viewed as the most basic piece of information. When a doctor examines a patient, the first task is to determine the clinical diagnosis. Then, the doctor can proceed with medical treatment based on that diagnosis. For hospital management, coded diagnoses are very important. Changes in the frequencies of diagnoses are important from the perspective of medical policy and budget making. However, it has been difficult to find investigations relating to the frequency distribution of diagnostic data. In this study, we found that the frequencies of clinical diagnoses have certain statistical features in common with natural languages.

As for natural languages, it is known that the frequencies of words follow what is called Zipf's law.[1–3] To apply this law, one calculates the frequencies of words within a given text. If all the words in the text are arranged in rank order, from most frequent to least frequent, an inverse power-law relation holds between the frequencies and the rank of the frequencies. The frequency of the word F_n, which has the rank n, is expressed by the following equation

$$F_n = \frac{A}{n^{\zeta}}$$

where A is constant. In this equation ζ is referred to as Zipf exponent, and in natural languages the exponent ζ was found to be close to one. In other words, log-log plots of the frequency versus rank show a linear relationship between these two variables. This relation is called Zipf's law. This law holds not only for languages, but also for many situations (e.g. the population of cities and their rankings).[4] In particular, it was recently discovered that DNA sequences have the same linguistic features.[5–12]

Our objectives were to determine whether the frequencies of clinical diagnoses follow Zipf's law and to investigate the influence of coding diagnoses on the distribution of frequencies.

2. METHODS

We analyzed data of the diagnoses stored within the database of Toyonaka Municipal hospital. Two data sets were analyzed separately; one is the set of diagnoses written on charts before October 1997, and the other is the set of diagnoses input into a computer after November 1997. The diagnoses before October 1997 were input into the computer by clerks, using the Japanese names on the charts of patients who had received medical advice at least once within the one-year period before October 1997. These data were input as Japanese text without code. Therefore, several names that indicate the same disease may be entered as different names due to differences in writing. Moreover, there is the possibility of errors due to incorrect input or typos on a chart, and in these cases, we input this erroneous name as a new different name. We think that the probability of making the same error twice is low, and in this case, the frequency of that diagnosis only amounts to one. A considerable number of the diagnoses showing a frequency of one is thought to be obtained by such errors and so we excluded all diagnoses showing a frequency of one from our analysis. The diagnoses were written on charts between July 1953 and October 1997. The total number of patients was 39 212 and the total number of diagnoses was 101 760, so the average was 2.6 diagnostic names per patient. Henceforth, we call this group the "freely written group."

The second group of data was collected between November 1997 and October 1998, and these data were directly input by doctors during this period. These data were basically input in the form of name codes that are registered as master data in the computer. Doctors input the Japanese name itself only when they were unable to find a corresponding name code. Entered name codes consist of six characters and the first four characters are the ICD10 codes (International Classification of Diseases 10th Revision). The last two characters are Toyonaka's own additional coding characters. With this additional coding, we can provide a more detailed categorization than ICD10 code alone. A correspondence map between this code and the Japanese disease name is registered as master data in the computer. By choosing a Japanese name on the computer, a doctor can input data with its code. Both code and the Japanese name are entered in the database file as the clinical diagnosis. In the present study, we used only the name codes for analysis. We did not use data without codes. The total number of patients was 47 964 and the total number of diagnoses was 137 933, so the average was 2.9 diagnoses per patient. Data that had no diagnosis codes comprised 3.3% of all data, and were excluded from further analysis. Henceforth, we call these data the "coded group."

We analyzed all data from all patients (in- and out-patients). With regard to a particular patient (and their initial diagnosis), the diagnosis (i.e. the name of the disease) may change after the first medical examination based on more detailed medical tests. However, we counted all diagnoses as distinct data.

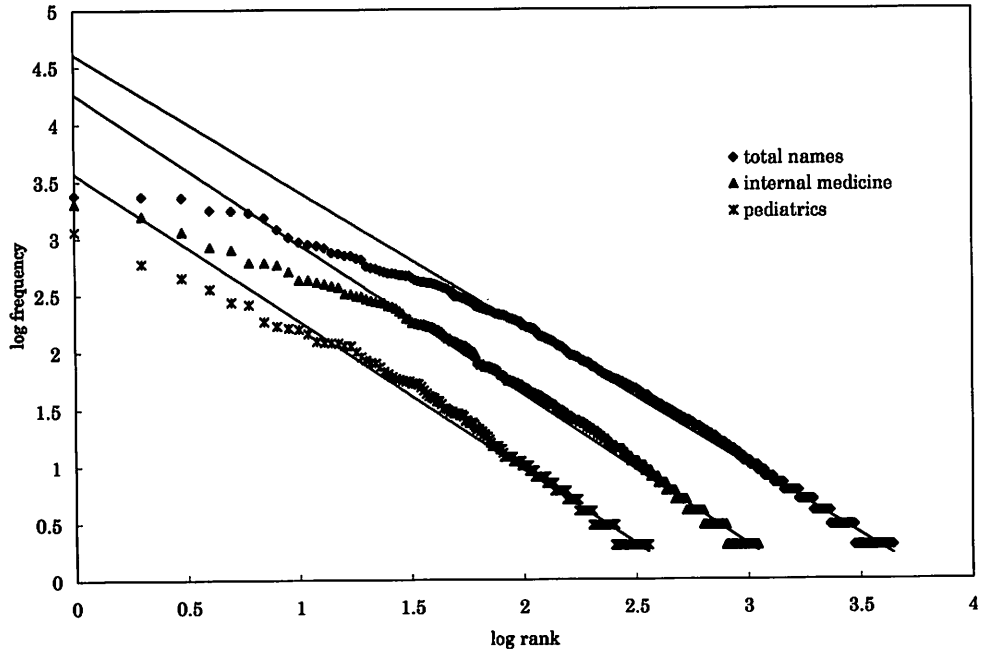
We evaluated the frequencies of clinical diagnoses for each group and the rank of each diagnosis (the most frequent diagnosis has the rank of one, the second most frequent diagnosis has the rank of two, and so on). Moreover, we plotted a logarithm of the frequency versus a logarithm of the rank (Zipf plot[7]). In the freely written group, we defined two names (Japanese names) as the same if and only if all characters contained in each diagnosis were completely identical (e.g. "Gastric Cancer" and "Cancer of Stomach" are regarded as different diagnosis). Therefore, we sometimes counted two diagnoses indicating the same disease as different ones only due to differences in the description.

In order to investigate the details of the distribution, we classified each data set into subgroups according to departments from which patients received medical advice and attention. Furthermore, with respect to the coded group, we also classified these data into subgroups according to doctors. These subgroups were also analyzed using a Zipf plot.

3. RESULT

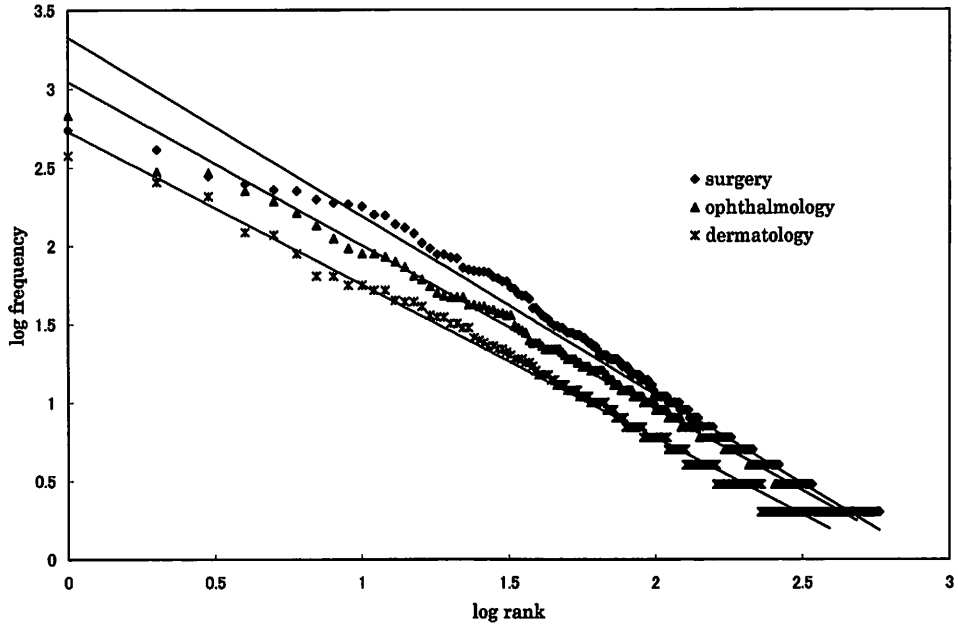
3.1 Freely Written Group

Figure 1 shows the result of the Zipf plot in the freely written group. In this figure (and these plots, in general), the x-axis denotes the logarithm of the rank of frequencies and the y-axis denotes the logarithm of frequencies. Moreover, regression lines for each data set are also expressed in the figures. For all departments, the data are fitted well by regression lines. These results mean that for all departments, the data follow Zipf's law. However, at high frequencies (from rank 1 to 10, i.e. from 0 to 1 on the x-axis) the data deviate less from the regression lines. These deviations are large in the "total diagnoses" group and in the internal medicine group. Causes of these deviations, especially among the total diagnoses and the internal medicine groups, will be discussed in the next section. Table 1 shows slopes of the regression lines for some



(a)

Fig. 1 Zipf plots of the freely written group. A logarithm of the frequency vs. a logarithm of the rank is plotted on the figures. The straight lines denote regression lines of data: (a) total diagnoses, internal medicine and pediatrics, and (b) Surgery, ophthalmology and dermatology.



(b)

Fig. 1 (*Continued*)

Table 1 Results of regression analysis.

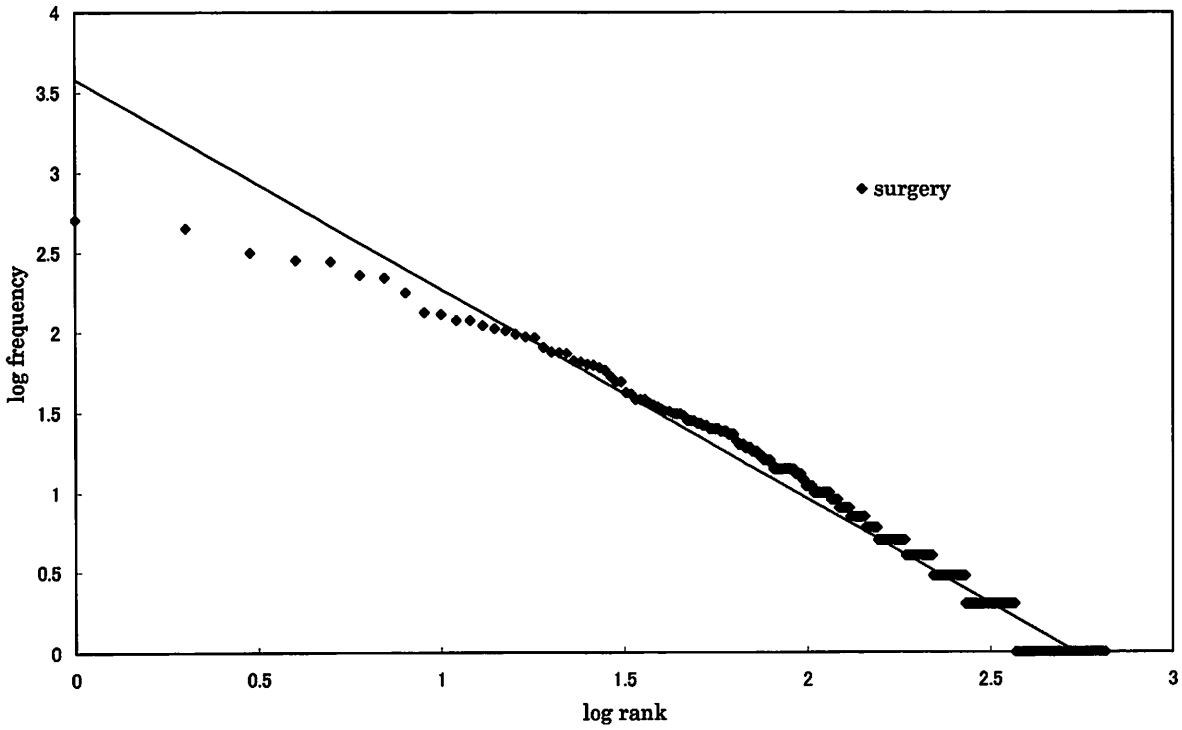
	Free written group			Coded group		
	ζ	r^2	N	ζ	r^2	N
Total	1.20 ± 0.002	0.99	93 253	1.64 ± 0.007	0.95	132 621
Internal medicine	1.32 ± 0.006	0.98	29 640	1.58 ± 0.009	0.97	29 634
Pediatrics	1.30 ± 0.009	0.98	8120	1.43 ± 0.009	0.97	12 164
Surgery	1.14 ± 0.007	0.98	7946	1.31 ± 0.009	0.97	7278
Neurosurgery	0.96 ± 0.008	0.98	1995	1.07 ± 0.014	0.96	1358
Ophthalmology	1.04 ± 0.005	0.99	5907	1.65 ± 0.028	0.94	6440
Orthopedics	0.83 ± 0.004	0.98	5770	1.58 ± 0.020	0.94	10 350
Cardiac surgery	0.97 ± 0.023	0.94	831	0.97 ± 0.014	0.96	791
Dermatology	0.98 ± 0.006	0.98	3623	1.41 ± 0.012	0.96	6950
Urology	1.25 ± 0.009	0.99	4289	1.50 ± 0.015	0.97	5227

Note: The result of regression analysis in the free written group and coded group. The Zipf exponent ζ and the coefficient of determination r^2 are obtained by linear regression analyses for the data consisting of logarithms of the frequency and the rank of diagnosis. N denotes the number of data of each subgroup.

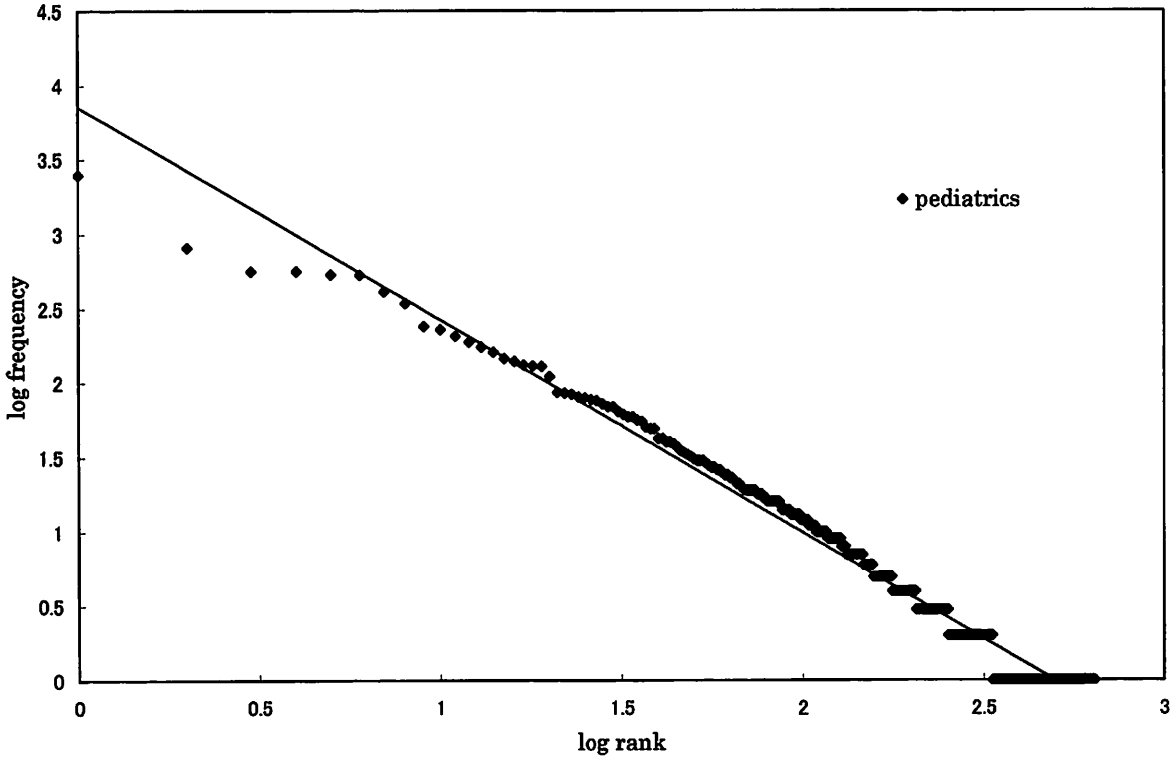
departments and for the total set. In the freely written group, these slopes are close to one.

3.2 Coded Group

For all groups except the total diagnoses and the internal medicine groups, data are fitted well by Zipf's law, especially at the low frequency regions [Figs. 2(a)–(c)]. As shown in Fig. 2(d), subgroups corresponding to doctors are also follow Zipf's law. However, for internal medicine and total diagnoses, data points deviate considerably from the predicted straight lines (Fig. 3). The deviations from straight lines at high frequencies are larger than those in the freely written group. These phenomena will be described in detail later.

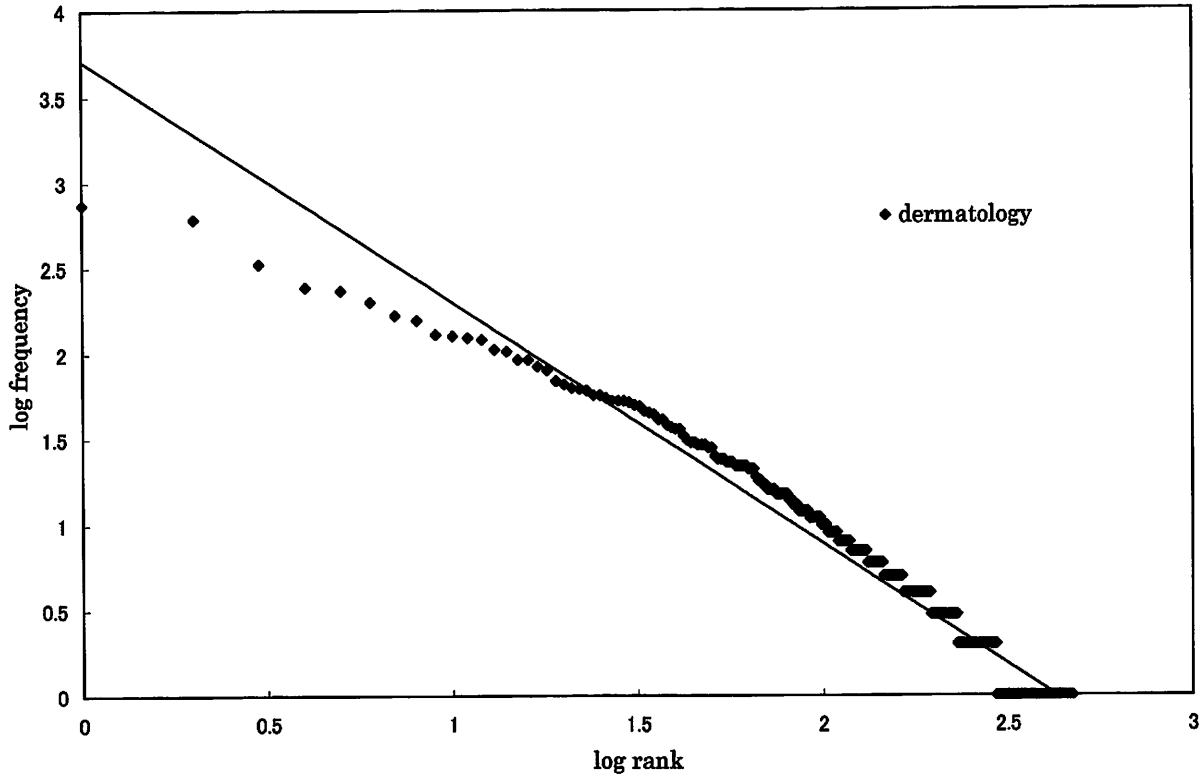


(a)

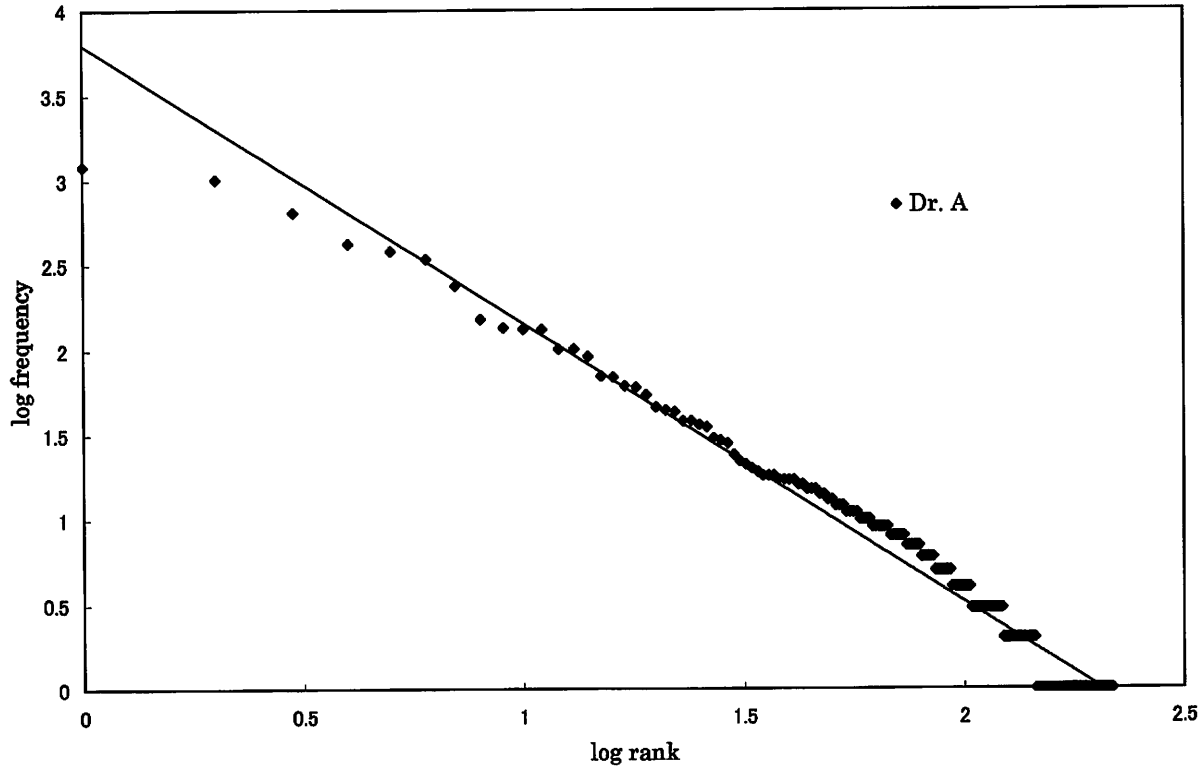


(b)

Fig. 2 Zipf plots of the coded group. The straight lines denote regression lines of data. (a), (b) and (c) denote surgery, pediatrics and dermatology, respectively, and (d) denotes data of doctor A.



(c)



(d)

Fig. 2 (*Continued*)

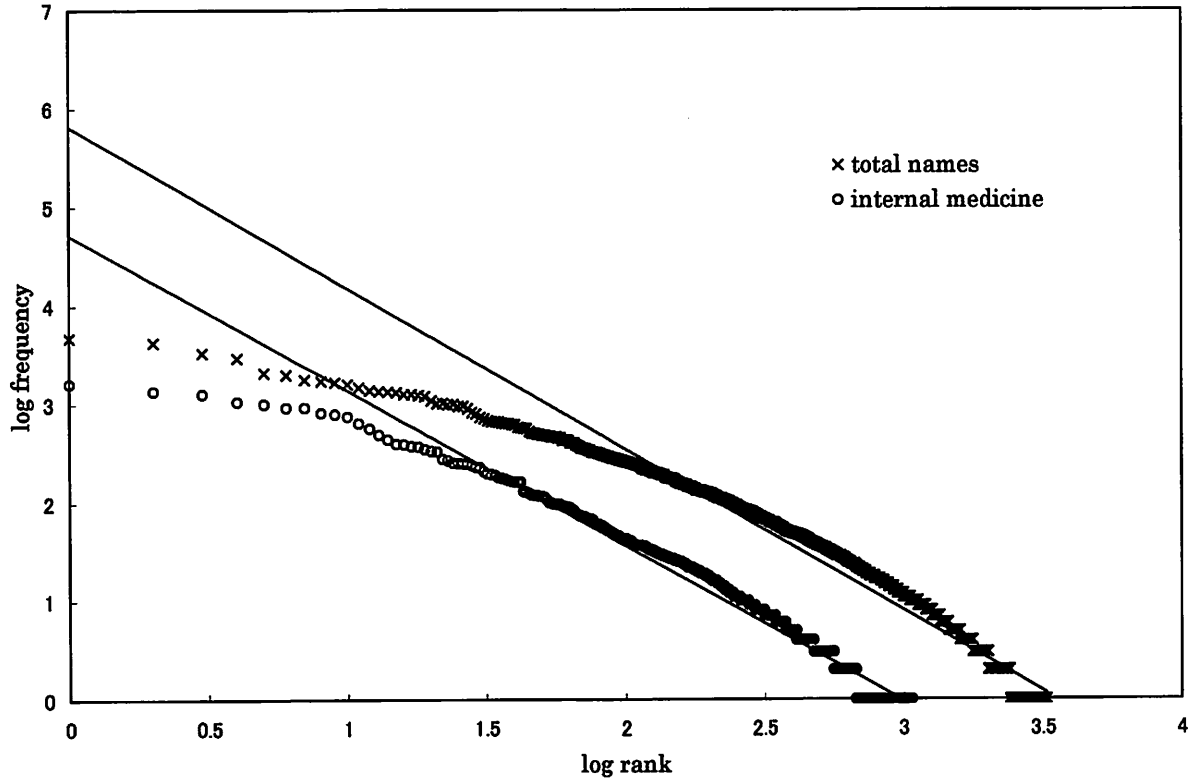


Fig. 3 Zipf plots of total diagnoses and internal medicine in the coded group. The data show large deviations from the regression lines, which are denoted by the straight line in the figure.

4. DISCUSSION

4.1 The Union of Zipf Sets Causes the Deviation

We indicate that in almost all groups, the distribution of clinical diagnoses follows Zipf's law. However, in analyzing the details of the results, we can recognize the deviation from Zipf's law at high frequencies.

We begin by considering the result obtained from the present study. A diagnostic set of a certain department (e.g. surgery, pediatrics, etc.) is the union of diagnostic sets of all doctors who belong to this department. From the present results, the sets of doctors are Zipf sets (a Zipf set indicates a set that follows Zipf's law), and the union of these sets (i.e. the set of the department) is also a Zipf set. This fact suggests that the union of several Zipf sets is also a Zipf set. Similarly, the total set of the hospital is the union of the sets of all departments and the total set is also a Zipf set. Moreover, the deviation at high frequencies of the total set is large compared to those of the sets of individual departments. This observation suggests that the union of Zipf sets has the large deviation at high frequencies compared to the original Zipf sets.

We next consider a simple mathematical example.

Example. Suppose two sets of equal size, which have no common diagnosis, and follow Zipf's law with a Zipf exponent of 1. For instance, data sets from ophthalmology and urology satisfy this condition. We suppose the numbers of diagnoses are both N. Combining the two sets, the number of diagnoses amounts to $2N$. If two sets have the same distribution and do not have any common diagnosis, the diagnosis with rank n in the source sets will have the rank $2n(=K)$ or the rank $2n-1$ in the unified set. This frequency is A/n (A is constant). Thus, the frequency Y_K for the diagnosis with rank K in the new set is expressed in terms of the following equation: $Y_K = Y_{2n} = A/n = 2A/2n = 2A/K = Y_{2n-1} = Y_{K-1}$. Therefore, for a large n, Y_K asymptotically approaches Zipf's law, and has the same Zipf exponent as the source sets. However, at high frequencies (in other words, where K

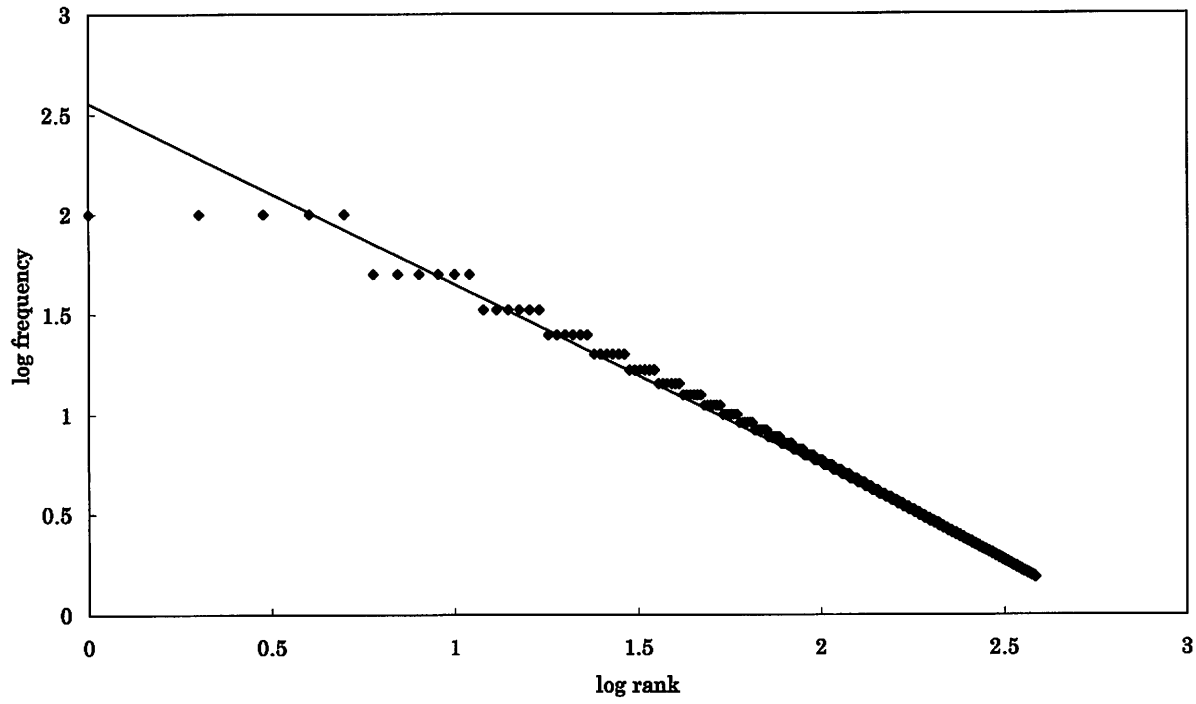


Fig. 4 Zipf plot of the union of five Zipf sets. The Zipf plot of the union of five sets, each of which follow Zipf's law and has no common diagnosis. Figure shows that this union asymptotically follows Zipf's law, but at high frequencies, data deviate from the regression line.

is small), it deviates from Zipf's law. Figure 4 shows the Zipf plot of the union of five sets with no common diagnoses. At low frequencies, we can see that the data are well fitted by a straight line. However, at high frequencies, data deviate from the expected line with the data points falling below the line.

This simple example indicates that the union of several sets, which follow Zipf's law and have no common diagnosis, is asymptotically fitted by Zipf's law. However, at high frequencies, such sets will deviate from Zipf's law.

This theoretical consideration, as well as the results obtained from the present data, indicates that the union of several Zipf sets is one of the causes of the deviation from Zipf's law at high frequencies. Moreover, this suggests that the sets of diagnoses for this analysis (total set of diagnoses, sets of each department, and sets of each doctor) consist of the union of some smaller subgroups, each of which follows Zipf's law.

In the analysis of the coded group, the deviations from straight lines at high frequencies are larger than those in the freely written group, especially in internal medicine and the total set of diagnoses. This suggests some other cause in addition to the union of Zipf sets. A possible cause is the saturation of diagnoses, which indicates the situation where the number of unused diagnoses decreases as a result of using too many names. In the coded group, the number of usable diagnoses is limited to the number of codes that are registered. The number of master data being used at this time is about 12 000. If the number of names used for patients increases to approaching the upper limit in the master data, it is possible that the saturation of diagnoses causes a deviation from Zipf's law. In the case of the total set of diagnoses, the number of diagnoses used is over 3200, so it is possible that the influence of the saturation of diagnoses is shown.

4.2 Giving a Clinical Diagnosis Depends on Context

According to this study, we found that the set of clinical diagnoses follows Zipf's law, or that the sets of clinical diagnoses consist of the union of several sets that follow Zipf's law. It is surprising that a common order (i.e. Zipf's law) can be seen between several types of diagnostic set (e.g. sets of department, sets of doctors, etc.). Conventionally, a

diagnosis was considered to be an objective label that was only dependent on each patient. Moreover, a disease (diagnosis) was thought to objectively exist within the population. The patient group of a given hospital is considered to be a sample of the population with a given diagnosis frequency. Thus, the diagnostic frequency of that patient group was assumed to be dependent on the population and the sampling method employed. According to this understanding, it would be difficult to assume that Zipf's law would apply as a common law to all sample groups (by department and doctor), including the total diagnostic group, that are sampled from the population by different methods.

Zipf's law appears to apply to the relationship between word frequencies of natural languages, city population, and the rankings that are derived from them. This law has been highlighted in relation to nonlinear dynamics that are represented by Fractal Mathematics and Chaos Theory.[1,13] As implied by its name, nonlinear dynamics studies systems (such as those in physiology) in which output is not proportional to input.[14] It is also a study of systems, which are described by several variables which interact nonlinearly. A Zipf's distribution is suggested as a distribution created from such nonlinear interaction. While the explanation of Zipf's distribution still remains unclear, the inter-variable correlation (interaction) over a long period is undoubtedly an important factor.[15] The probability of word appearance in a natural language, for example, is not independent and would be affected by the context. This is considered a necessary condition to create a Zipf's distribution. In other words, it is assumed that the word is context-dependent.

When applying this concept to a diagnosis, it indicates that a patient diagnosis may be affected by the diagnosis of a previous patient. This suggestion is surprising since it contradicts conventional understanding that a diagnosis was dependent only on the patient.

We will further review the process involved for making a diagnosis. Not all clinical diagnoses are independent of each other. In fact, some diagnoses are strongly related one another. For example, a patient who shows symptoms of a common cold, may be given a diagnosis of one of following diseases: common cold, upper respiratory infection, acute rhinitis, acute pharyngitis or acute bronchitis. These names are easy to use interchangeably (or to diagnose in error). In this sense, these terms are closely related to each other; in other words, the "distance" between them is short. However, a femoral fracture and an upper respiratory infection would not be confused, so we think that these two diagnoses are unrelated and the distance between them is great.

When diagnosing a patient, a doctor must choose a diagnosis that most closely resembles the symptoms. For instance, consider a situation where several close diagnoses fit the symptoms of the patient, but there is no absolute standard for selecting one of these due to the vagueness that each diagnosis has. In such cases, the selection of the diagnosis depends on the doctor's free choice. This situation is similar to a situation in which we select one from many possible words when we write a composition. In natural language, a word is selected based on the context. Similarly, free selection of a given diagnosis is affected by many factors. These, for example, include the diagnosis of a patient previously seen with similar symptoms, a recent diagnosis frequency of that hospital, the doctor's skill, and so on. As a result, determining a diagnosis is not only dependent on the patient's condition, but also on the conditions of the doctor, previous patients and the hospital, as well. It is assumed that this effect creates an interdiagnostic correlation and this correlation may satisfy a necessary condition to create a Zipf's distribution.

A hospital is a system that consists of complicated interactions between doctors, patients, nurses, etc. We think that these complicated interactions yield a new order called Zipf's law and this new order cannot be predicted by any study of each element (doctors, patients, nurses, etc.). This means that it is important to regard a hospital system as a "complex system" as a whole.[1]

Lastly, let's think about the effects of Zipf's distribution on medical service. In a set that follows Zipf's law, the relative frequency of a certain element (i.e. the frequency of a certain element/the total frequency) depends on the size of the set.[16–18] If, for example, the Zipf exponent is 1, the relative frequency of a rank 1 diagnosis is 0.28, when the group size is 100. However, the relative frequency of the same diagnosis would be 0.13 when the size is 10 000. Moreover, when the Zipf exponent is 1, the relative frequency tends to get closer to zero in proportion to the increase in group size. This is quite surprising and unintelligible considering that the group is a sample from the population. According to conventional rationales, the relative frequency of a diagnosis was thought to

be its relative frequency in a group sampled from the assumed virtual population. In this case, the variation in relative frequency was thought to be a random fluctuation according to the sampling procedures, and therefore, was assumed to be a normal distribution around a given relative frequency. The understanding was that the true relative frequency was sought in the form of a mean relative frequency after repeated sampling. However, it is not possible to define the true diagnostic relative frequency according to the above understanding in the case of a Zipf distribution where the relative frequency varies according to the group size. This may show that the measure of "the relative frequency of a diagnosis" obtained by extension of the relative frequency of a sample does not originally exist. This is similar to the situation in which the idea "a curve without length or a plane without area" arises from the discovery of "Fractal."[19]

4.3 Many Medical Indices also Follow Zipf's Law

Though we do not present the data at this time, in addition to the clinical diagnoses, medical indices such as average length of hospital stay, frequencies of medical treatments expressed in terms of ICD9-CM (International Classification of Disease 9th Revision, Clinical Modification) and medical fees, also follow Zipf's law. These facts suggest that not only the clinical diagnoses but also the medical structure itself is based on complex interactions between patients and a medical system that includes a medical team, medical facilities and medical equipment, and all of these interactions yield Zipf's law. A more detailed study of the healthcare delivery structure on the basis of the theory of complex systems is required.

5. CONCLUSION

It was proven that the diagnostic sets based on the doctor's diagnoses followed Zipf's law. This indicates that the diagnostic set is a set interactively created by the doctor, patient and hospital, and it follows a certain order called Zipf's law. This fact indicates that the medical system, consisting of doctor-patient interaction, is a so-called complex system. Furthermore, the indication that diagnostic sets observe Zipf's law may possibly have major effects on changing the conventional concept of diagnostic frequency rate.

ACKNOWLEDGMENTS

The authors would like to thank Professor I. Tsuda (Hokkaido University) H. Takayasu (Sony Computer Science Laboratories, Inc.) and Y. Hirota for their useful discussions.

REFERENCES

1. J. L. Casti, "Bell Curves and Monkey Languages," *Complexity* 1(1), 12–15 (1995).
2. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge MA, Addison-Wesley, 1949).
3. D. R. Ridley and E. A. Gonzales, "Zipf's Law Extended to Small Samples of Adult Speech," *Percept. Mot. Skills* **79**, 153–154 (1994).
4. A. S. Iberall, H. Soodak and F. Hasslerm, "A Field and Circuit Thermodynamics for Integrative Physiology. II. Power and Communicational Spectroscopy in Biology," *Am. J. Physiol.* **234**, R3–19 (1978).
5. R. F. Voss, "Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences," *Phys. Rev. Lett.* **68**, 3805–3808 (1992).
6. R. N. Mantegna and S. V. Buldyrev, A. L. Goldberger, et al. "Linguistic Features of Noncoding DNA Sequences," *Phys. Rev. Lett.* **73**, 3169–3172 (1994).
7. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, el al. "Systematic Analysis of Coding and Noncoding DNA Sequences Using Methods of Statistical Linguistics," *Phys. Rev.* **E52**, 2939–2950 (1995).
8. J. D. Burgos and P. Moreno-Tovar, "Zipf-Scaling Behavior in the Immune System," *Biosystems* **39**, 227–232 (1996).
9. B. J. Strait and T. G. Dewey, "The Shannon Information Entropy of Protein Sequences," *Biophys. J.* **71**, 148–155 (1996).
10. J. D. Burgos, "Fractal Representation of the Immune B Cell Repertoire," *Biosystems* **39**, 19–24 (1996).
11. C. A. Chatzidimitriou-Dreismann, R. M. Streffer, D. Larhammar, et al., "Lack of Biological Significance in the 'Linguistic Features' of Noncoding DNA — A Quantitative Analysis," *Nucleic Acids Res* **24**(9), 1676–1681 (1996).
12. A. A. Tsonis, J. B. Elsner and P. A. Tsonis. "Is DNA a Language?" *J. Theor. Biol.* **184**(1), 25–29 (1997).
13. J. S. Nicolis and I. Tsuda, "On the Parallel Between Zipf's Law and 1/f Processes in Chaotic Systems Possessing Coexisting Attractors," *Prog. Theor. Phys.* **82**(2), 254–273 (1989).
14. L. A. Lipsitz and A. L. Goldberger. "Loss of 'Complexity' and Aging Potential Applications of Fractals and Chaos Theory to Senescence," *JAMA* **267**, 1806–1808 (1992).

15. P. Bak, C. Tang and K. Wisenfeld, "Self-Organized Criticality," *Phys. Rev.* **A38**, 364–374 (1988).
16. B. B. Mandelbrot, "The Stable Paretian Income Distribution When the Apparent Exponent is Near Zero," *Int. Econ. Rev.* **4**, 111–115 (1963).
17. B. B. Mandelbrot, "The Variation of Certain Speculative Stock Prices," *J. Bus.* **36**, 394–419 (1963).
18. B. B. Mandelbrot, "New Methods in Statistical Economics," *J. Polit. Econ.* **71**, 421–440 (1963).
19. B. B. Mandelbrot, *The Fractal Geometry of Nature* (New York, Freeman, 1983).