J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, Eiji Aramaki

Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{kito, hironagai, taro-o, wakamiya, iwao, aramaki}@is.naist.jp

Abstract

Medical texts such as electronic health records are necessary for medical AI development. Nevertheless, it is difficult to use data directly because medical texts are written mostly in natural language, requiring natural language processing (NLP) for medical texts. To boost the fundamental accuracy of Medical NLP, a high coverage dictionary is required, especially one that fills the gap separating standard medical names and real clinical words. This study developed a Japanese disease name dictionary called "J-MeDic" to fill this gap. The names that comprise the dictionary were collected from approximately 45,000 manually annotated real clinical case reports. We allocated the standard disease code (ICD-10) to them with manual, semi-automatic, or automatic methods, in accordance with its frequency. The J-MeDic covers 7,683 concepts (in ICD-10) and 51,784 written forms. Among the names covered by J-MeDic, 55.3% (6,391/11,562) were covered by SDNs; 44.7% (5,171/11,562) were covered by names added from the CR corpus. Among them, 8.4% (436/5,171) were basically coded by humans), and 91.6% (4,735/5,171) were basically coded automatically. We investigated the coverage of this resource using discharge summaries from a hospital; 66.2% of the names are matched with the entries, revealing the practical feasibility of our dictionary.

Keywords: Medical NLP, Case reports, Discharge summary, Named entity, ICD-10

1. Introduction

Medical data are fundamentally important resources for the development of medical AI and information extraction tools. Among various data, Electronic Health Records (EHR) are a promising resource because they include detailed information about a patient and diagnosis processes. Nevertheless, it is difficult to extract information from EHR because several expressions refer to the same concept. Orthographical variations present particular difficulty, especially for the Japanese language, in which characters of at least five kinds are used: Hiragana, Katakana, Kanji, Latin alphabet, and Arabic Numerals. In addition to orthographic difficulties, variations of expression for the same concept delivered by other reasons such as abbreviations are included in medical texts produced at clinics, hospitals, and other medical institutes. These variations present obstacles that are encountered when developing medical AI or information extraction tools, although several studies have been undertaken to solve these and related difficulties by developing or assisting automatic coding systems (Fabry et al., 2003; Yamada et al., 2010; Bouchet et al., 1998).

Actually, several medical resources exist for non-Japanese languages, such as the International Classification of Diseases (ICD; World Health Organization, 2004) [1], Medical Subject Headings (MeSH; Lipscomb, 2000), and Systematized Nomenclature of Medicine Clinical Terms (SNOMED-C[2]; Benson, 2012). SNOMED-CT, the largest, includes approximately 308,000 concepts and 777,000 expressions, and officially supports English and Spanish. Also, projects are translating SNOMED-CT into other languages (Abdoune et al., 2011; Zhu et al., 2012).

Currently, medical language resources for Japanese language are smaller than those for other major languages such as English. The ICD10-based Standard Disease-Code Master (SDCM; Hatano and Ohe, 2003) [3] is the most widely used resource in Japan. The current version of SDCM covers approximately 24,000 disease names. Each name has a corresponding ICD-10 code.

This study was conducted to solve such a problem, developing a dictionary of disease names that appears in medical texts. First, we collected over 45,000 medical case reports from the Japanese Society of Internal Medicine. After we annotated disease expressions in case reports automatically using a named entity recognition (NER) tool to reduce the related work, 13 annotators amended them manually. Next, we split the disease list into three sub-lists: high-frequency, middle-frequency, and low-frequency parts. For the high-frequency part, three human coders manually allocated codes. The middle-frequency part was divided into three subparts: and each of the three coders coded each subpart. For the low-frequency part, we automatically added codes using a classifier trained with the high-frequency part.

Characteristics of the dictionary we developed, the Japanese Medical Dictionary (J-MeDic), are explained below.

- Entries were collected by the Japanese Society of International Medicine and were validated using data obtained from the University of Tokyo Hospital.
- Wide varieties of the expression for an illness are included. The average number of variants for one concept (ICD-10) is 6.74.
- Each entry has information about the corresponding ICD-10. Via ICD-10, Japanese names can be translated into various languages in which ICD-10 is available.

[1] http://www.who.int/classifications/icd/en/

[2] http://www.snomed.org/snomed-ct/

[3] http://www.dis.h.u-tokyo.ac.jp/byomei/index.html

2. Materials

To construct J-MeDic, we used the following three materials for different purposes: ICD-10 and ICD-10-based Standard Disease Code Master for the basis of classification; CR corpus (case reports) for extracting new vocabularies that represent disease names; and HDS corpus (hospital discharge summaries obtained from the University of Tokyo Hospital) for validation of J-MeDic. Both corpora consist of electrical data.

2.1 ICD-10 and ICD10-based Standard Disease Code Master

ICD-10 is the diagnostic classification standard for clinical and research purposes. It is also the international standard for reporting diseases and health conditions. It classifies diseases, disorders, injuries, and other related health conditions (hereinafter, "diseases") in a comprehensive and hierarchical fashion. The ICD code comprises an alphabet and 2–4 digits. The first character of an ICD code is an alphabet called an *axis*, which refers to the kind of disease followed by digits referring to a detailed site. For example, in "C341", the first character "C" refers to "malignant neoplasms". The following two digits "34" refer to "malignant neoplasm of bronchus and lung" together with "C". The last digit "1" means "upper lobe". To match Japanese language expressions with ICD-10 code, we used SDCM, which provides an interface on the website for retrieval of ICD code from natural language and vice versa.

2.2 CR Corpus

To collect names for diseases, we used the CR corpus, an annotated corpus of 44,761 case reports. The Japanese Society of International Medicine provided the reports. After annotating the disease names in the corpus, we extracted them to collect entry candidates for J-MeDic.

2.3 HDS Corpus

The HDS corpus holds discharge summaries from 291,641 patients hospitalized at the University of Tokyo Hospital, Japan, between 2004 and 2016. These summaries, which include brief descriptions of diagnoses, clinical outcomes, comorbidities and observations on admission, and post-admission clinical course, were used for J-MeDic validation: We counted the frequency of the names included in J-MeDic to confirm the extent to which J-MeDic covers the names for diseases of other real clinical texts.

3. Methods

3.1 Structure of J-MeDic

A record in J-MeDic has the following fields:

Name: The expression collected from the corpus, presumed to be used as dictionary entries

ICD code: The ICD code allocated to the name

Standard disease name: The standardized disease name (SDN, disease names collected and standardized in SDCM) allocated to the name

Reliability level: Level of reliability (S, A, B, C)

Kana: Pronunciation of the name, written in Hiragana characters

Site and symptom type: The site at which the disease happens and the type of symptom

Frequency JSIM: Frequency of the name in the data from the Japanese Society of Internal Medicine (i.e. CR corpus)

Frequency UTH: Frequency of the name in the data from the University of Tokyo Hospital (HDS corpus)

For example, Table 1 presents the record of the entry "糖尿病" (diabetes).

Field	Value
Name	糖尿病 (diabetes)
ICD code	E14
Standard disease name	糖尿病 (diabetes)
Reliability level	S
Kana	とうにょうびょう
Site and symptom type	region=膵臓/type=その他 (region=pancreas/type=other)
Frequency JSIM	61,572
Frequency UTH	5,645

Table 1: Record of *diabetes*. English translation added (enclosed in parentheses).

3.2 Entire Process of J-MeDic Construction

J-MeDic was constructed with the following steps.

1. CR corpus annotation
2. Extracting disease names from annotated CR corpus and ICD coding
3. Reliability assessment of the coding
4. Merger with SDNs
5. Evaluation of representativeness

In step 1, we first annotated the words that represent diseases in the corpus automatically; then we manually modified it (Section 3.3). In step 2, candidate entries for J-Medic were extracted from the annotated CR corpus. Then the coders allocated ICD-10 codes (and also the SDN that the allocated ICD accompanies) to these names (Section 3.4). After finding the reliability of the ICD code(s) of each entry (step 3, Section 3.5), SDNs were merged with them (step 4, Section 3.6). Finally, we verified the representativeness of J-Medic using of HDS corpus (step 5, Section 3.7).

3.3 CR Corpus Annotation

To reduce manual annotation work, we automatically annotated disease names that appear in the CR corpus using MedEX/J, which is a tool for analyzing Japanese clinical text. It supports identification and extraction of text strings of potential disease names. Based on conditional random fields, the module learns disease name labels from a disease name annotated corpus. Details of this module are available (Aramaki et al., 2017).
Subsequently, 13 annotators including non-medical workers modified the CR corpus. To construct a resource

for disease names, we set the following criteria and annotated the corpus.

(I) To avoid complexity caused by reference to a disease by split words, the syntactic category of the target was limited to (compound) nouns.

(II) To extract as many candidate disease names as possible, the lexical units were labeled if suspected.

Of those, (I) is important in cases where a disease is referred by a noun, a verb, or other peripheral words. The following three expressions refer to almost identical situations (coded as N289 in ICD-10 system), but which are grammatically different:

(a) **腎機能低下**が見られた (***renal function degeneracy*** *was observed*)

(b) **腎機能**が**低下**していた (***renal function*** *had* ***degenerated***)

(c) **腎機能**が高度に**障害されて**いた (***renal function*** *had severely* ***degenerated***)

In (a), the disease is referred to by a single compound noun (*renal function degeneracy*), whereas the disease is referred by separate words in (b) and (c) (*renal function* and *degenerated in* both cases). In the latter cases, it is difficult to delineate the exact boundary because of its syntactic complexity. Therefore, we limited the syntactic unit of the annotation target.

In addition, (II) is important to maximize the number of disease names that were not yet collected from other language resources. In other words, it is important to expand the vocabulary in J-MeDic. Although technical terms should be standardized, disease names in real case reports are expressed in abbreviated form or in a slightly modified way. In the step of manual annotation, we collected such variations to the greatest extent possible. Then inappropriate variations were excluded from the coding step.

3.4 Coding Procedures

3.4.1 Coders

In this study, three coders coded the data. All the coders had work experience as health care staff.

3.4.2 Collection of Disease names and Manual Coding

Three coders coded high-frequency and middle-frequency parts before automatic coding for the low-frequency part. For the high-frequency part, all coders coded the entire part, discussing it if needed. Each of the three coders coded different subparts for the middle-frequency part. Figure 1 presents a coding process summary.

From the annotated CR corpus, we extracted all the disease names appearing in the corpus. Then, the coders coded disease names in ICD-10. First, we searched the exact matches of the SDNs Master for all the disease names. If a disease name had an exact match, then it was allocated the corresponding ICD of the SDN.

Next, for names that had not been coded in the prior step, the coders searched the exact match of the transliteral variations. For example, a person name "Wegener" appears as it is (i.e. in Latin alphabet) or as various transliteration into Japanese characters such as "ウェゲナー" and "ウェジナー". Therefore, to code "Wegener 肉芽腫" (Wegener's granulomatosis), the coders sought an exact match of "ウェゲナー肉芽腫" and "ウェジナー肉芽腫".

After searching orthographical variants, the coders searched partial matches of the remainder of disease names to avoid extra modifiers. The coders tried queries that are created by omitting modifiers in the name. For example, "LQT2 型 QT 延長症候群" (LQT2 type long QT syndrome) does not match any SDN. In this case, the deletion of "LQT2 type" allows matching of "long QT syndrome". It is coded as I490 (Ventricular fibrillation and flutter). Furthermore, guessing from that "LQT2" refers to a kind of gene, "LQT2 type long QT syndrome" can be categorized as a subcategory of "inherited long QT syndrome", which is listed in corresponding standardized disease name to I490. Therefore, "LQT2 type long QT syndrome" was coded as I490, and standardized as "inherited long QT syndrome".

When multiple ICD codes correspond to a name, we allocated up to two codes. If more than two possible codes were found, then the name was excluded from the targets of ICD coding, and was coded as "-1". In case reports, multiple nouns that represent a disease often appear together to form a compound noun. For example, "脂肪肝合併 2 型糖尿病" (type 2 diabetes complicated with fatty liver) is divisible into "type 2 diabetes" and "fatty liver". Therefore, we allocated both codes to this name. There were also other cases in which multiple codes are allocated to a name: (i) the concept represented by the name is too vague and (ii) the interpretation of the name differs from coder to coder.

If no matched SDN was found after these steps explained

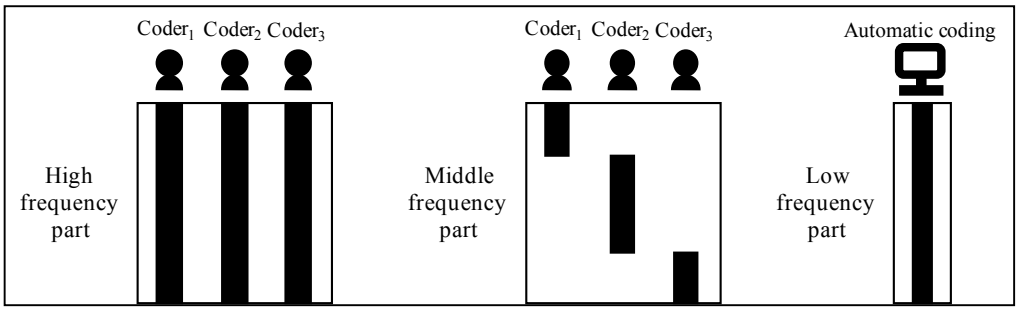

Figure 1: Summary of the coding. A black band represents that the data were coded by the worker indicated above.

above, then the name was coded as "-1".

3.4.3 Comparison and Discussion

Among extracted disease names from the corpus, all the three coders coded the names that appeared more than 29 times (hereinafter, *high-frequency names*). The names appeared more than 3 and fewer than 30 times (hereinafter, *low-frequency names*) divided into three parts and each part was coded by a coder. When a coder found the name difficult to code, the three coders discussed the coding.

The results of coding for frequent names sometimes varied among coders. The final codes for frequent names were decided using the following criteria.

I. When all coders judged a name as not a target, the name was coded as "-1".
II. When all coders allocated the same ICD code to a name, the code was adopted as the final version.
III. When coders allocated a name to different codes, they discussed it and chose a code.

3.5 Reliability Assessment

Based on the coding results, we decided the reliability level of the names depending on how the code was decided. The levels were defined as explained below.

S: Matched with a SDN

A: Three coders allocated the code

B: Coded by one coder, or two coders if discussed

C: Automatically coded, pended, or excluded from the target

3.6 Merger with SDNs

To expand J-MeDic, we merged SDNs with disease names extracted from CR corpus. All the SDN were included into J-MeDic with reliability level "S". Regarding disease names that are extracted from HDS corpus and which are not in annotated CR corpus nor SDN. Disease names collected from HDS corpus were coded automatically.

3.7 Evaluation of Representativeness

To evaluate the representativeness of the entries in J-MeDic, we calculated the coverage of the disease names that appear more than nine times in HDS corpus. The disease names on HDS corpus were extracted using the natural language processing module MedEX/J.

4. Results

4.1 Overview: The J-MeDic size

The J-MeDic covers 7,683 ICD-10 concepts (when a disease name was allocated two ICD codes, the pair was counted as one concept) and 51,784 written forms. Among the written forms, 25,365 (49.0%) forms were not contained in SDCM.

4.2 Matching with SDN

From CR corpus, 30,923 disease names were extracted. Among them, 5,558 names were matched exactly with a SDN and coded as the corresponding ICD-10 of the SDN.

4.3 High-frequency Names

Except for exact-matched disease names with SDN, 804 high-frequency names were found. Table 2 shows the result of coding for high-frequency names by the three coders.

Coding category	# of names (n = 804)
Pended	165
Same code allocated	467
Differently coded	172

Table 2: Primary result of the coding on high-frequency names

We also calculated the inter-rater agreement among the disease names that are not excluded from the target (Table 3). Each coder is represented by c_i.

coder pair	ratio (%)
c_1, c_2, and c_3 allocated the same code	73.1 (467/639)
Only c_1 and c_2 allocated the same code	7.5 (48/639)
Only c_2 and c_3 allocated the same code	7.8 (50/639)
Only c_2 and c_3 allocated the same code	7.5 (48/639)

Table 3: Agreement between coders

4.4 Middle-frequency Names

Except for exact-matched disease names with SDN, 5,319 middle-frequency names were found. Table 4 presents results of coding on low-frequency names.

Coding category	# of names (n = 5,319)
Coded by one coder	4,710
Decided after discussion	3
Pended	606

Table 4: Result of the coding on high-frequency names

4.5 Low-frequency Names

After annotating high-frequency and middle-frequency names, low-frequency names were annotated automatically using backward matching. For each low-frequency name, we found the longest backward match with the high-frequency and middle-frequency names. Then we allocated its ICD (or ICDs, if it has two codes) to the name. Considering the morphological structure of disease names (i.e. most of the head of a compound noun occupies the latter part), we used backward matching.

4.6 Reliability level

Table 5 shows the counts of the reliability level.

Reliability level	**# of names (n = 51,784)**
S	26,419
A	528
B	4,808
C	20,029

Table 5: Reliability level

4.7 Coverage

In the HDS corpus, 17,469 disease names appeared more than nine times. J-MeDic covers 66.2% (11,562/17,469) of these names. Among the names covered by J-MeDic, 55.3% (6,391/11,562) were covered by SDNs; 44.7% (5,171/11,562) were covered by names added from the CR corpus. Among them, 8.4% (436/5,171) were entries

with reliability level A or B (i.e. basically coded by humans), and 91.6% (4,735/5,171) were entries with reliability level C (i.e. basically coded automatically).

5. Discussion

5.1 Extension of the Resource

As described in Section 5.1, J-MeDic contains 51,784 new written forms; 49.0% of those were newly incorporated. However, 44.7% of the disease names that are covered by J-MeDic were newly incorporated written forms. This result can be regarded as indicating that J-MeDic increased the number of the disease names included in a language resource by about 90%.

However, J-MeDic also has limitations. Among newly incorporated disease names that appeared in the HDS corpus, only 8.4% of the names were reliability level A or B, although 21.0% (5,336/25,365) of the disease names in J-MeDic are labeled as reliability level A or B. Because this ratio can differ depending on the corpus, it does not mean directly that the coverage of disease names with reliable ICD code in J-MeDic is low. Therefore, J-MeDic mainly contributed to extension of the entry because disease names of reliability level C are useful to search disease names, although their ICD codes are not reliable. At the same time, J-MeDic can partly be used to detect the particular diseases listed in ICD-10 written in various forms.

5.2 Problem Stemming from ICD

Some difficulties arise stemming from the system of ICD-10. First, because the criteria of the classification in ICD-10 were not clear, coders sometimes had difficulties to search or to identify the ICD code that correspond to a disease name, especially in the Japanese version. For example, N40 (前立腺症, Hyperplasia of prostate) and N429 (前立腺障害, Disorder of prostate, unspecified) respectively correspond to similar disease names, but have different codes. Moreover, because the ICD code for a particular body part sometimes does not exist, the coders had to guess. Furthermore, some orthographical variations caused search difficulties.

5.3 Difference between Coders

Some limitations arose from different opinions among coders. One cause is that coders differently decided if a disease name is a target or not. For example, 転移 (*metastasis*) can be associated with 転移性腫瘍 (C80, Malignant neoplasm, without specification of site), and also 止血困難 (*difficulty in hemostasis*) can be associated with 出血 (R58, Haemorrhage, not elsewhere classified). However, these expressions are a clue to guessing the disease, but not the disease itself. We excluded such expressions from the target.

Another cause is that coders assigned some different codes to disease names. In such cases, the code was generally chosen by majority, but important minority opinions were considered and accepted sometimes. As Section 4.4 showed, up to two codes were allocated to one disease name when selecting only one was difficult.

6. Conclusion and Future Work

We developed J-MeDic, designed for automatic information extraction from medical texts. The newly incorporated words were collected from case reports to improve the coverage of orthographical variations appearing in unstructured texts. We believe that J-MeDic is useful in various fields: Not only can it be used to develop medical AI; it can also help medical workers to write documents using standardized medical terms.

Although we have extended language resources for disease names, numerous names labeled as reliability level D in J-MeDic were coded automatically and were not verified. In future work, further investigations will be necessary to improve the dictionary reliability.

7. Acknowledgements

This work was partly supported by JSPS KAKENHI (JP16H06395, JP16H06399), and by AMED (Grant Number: JP17lk1010019), Japan.

8. Bibliographical References

Abdoune, H., Merabti, T., Darmoni, S. J., and Joubert, M. (2011). Assisting the translation of the CORE subset of SNOMED CT into French. In *MIE* (Vol. 169, pp. 819-823).

Aramaki, E., Yano, K., and Wakamiya, S. (2017). MedEX/J: A One-scan simple and fast NLP tool for Japanese clinical texts. MEDINFO 2017: eHealth-enabled Health.

Benson, T. (2012). SNOMED CT Concept Model. In *Principles of Health Interoperability HL7 and SNOMED* (pp. 253-266). Springer London.

Bouchet, C., Bodenreider, O., and Kohler, F. (1998). Integration of the analytical and alphabetical ICD10 in a coding help system. Proposal of a theoretical model for the ICD representation. *Medinfo.* 9(1):176–179.

Fabry, P., Baud, R., Ruch, P., Le Beux, P., and Lovis, C. (2003). A frame-based representation of ICD-10. *Studies in Health Technology and Informatics*, 95:433–438.

Hatano, K. and Ohe, K. (2003). Information Retrieval System for Japanese Standard Disease-Code Master Using XML Web Service. *AMIA Annual Symposium Proceedings*, 859.

Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265–266.

World Health Organization. (2004). *International statistical classification of diseases and related health problems* (Vol. 1). World Health Organization.

Yamada, E., Aramaki, E., Imai, T., and Ohe, K. (2010). Internal structure of a disease name and its application for ICD coding. *Studies in Health Technology and Informatics*, 160(2):1010–1014.

Zhu, Y., Pan, H., Zhou, L., Zhao, W., Chen, A., Andersen, U., Pan, S., Tian, L., and Lei, J. (2012). Translation and Localization of SNOMED CT in China: A pilot study. *Artificial Intelligence in Medicine*, 54(2):147–149.