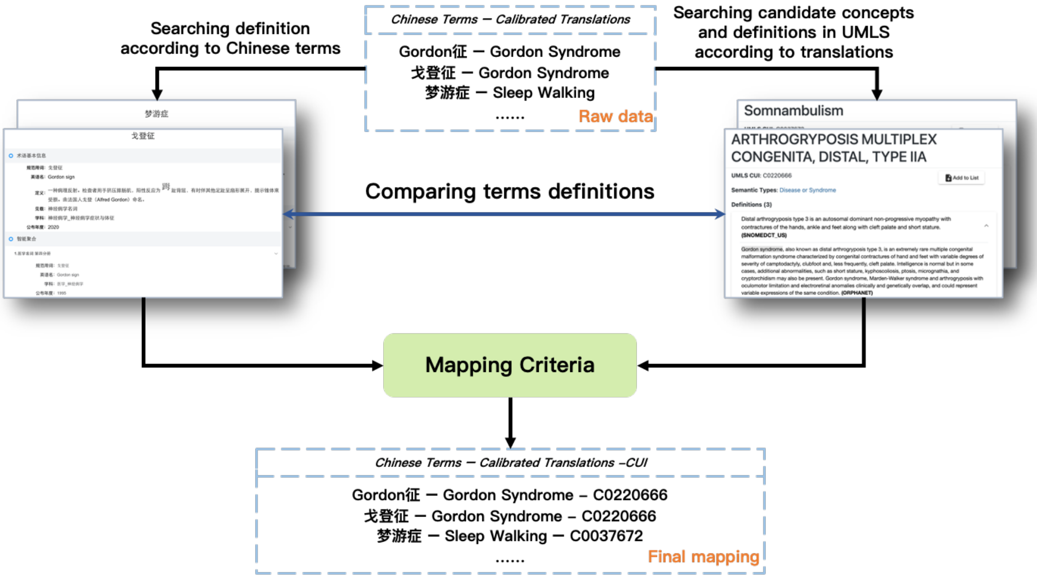
Annotation Guidelines for Mapping Chinese Medical Terms to the UMLS Concepts

1. Introduction

In this mapping annotation task, the main objective is to map Chinese medical terms which were extracted from real-world medical documents to the UMLS concepts. At First, the real-world medical documents including clinical EHRs, online health communities, case reports or other medical documents in real-world usage were fully annotated and reviewed in previous work [1], [2]. Then, all named entities were collected as the raw data of the dataset. The goal of annotators in the mapping task is to find the best UMLS concepts for the terms in the raw data. All mapped Chinese medical terms in the raw data are collected to form the dataset. To ensure the consensus, annotators are asked to follow the mapping steps and mapping criteria described in the following sections.

2. Mapping Steps

There are **FIVE** main steps for mapping Chinese medical terms to the UMLS. Figure S1 has briefly described the annotation process for mapping Chinese medical terms to the UMLS.


Searching definition according to Chinese terms
Chinese Terms – Calibrated Translations
Searching candidate concepts and definitions in UMLS according to translations
Raw data
Comparing terms definitions
Mapping Criteria
Chinese Terms – Calibrated Translations –CUI
Final mapping
Add to List

Gordon征 – Gordon Syndrome
戈登征 – Gordon Syndrome
梦游症 – Sleep Walking
……

梦游症

戈登征

术语基本信息

规范用词：戈登征
英语名：Gordon sign
定义：一种病理反射。检查者用手挤压腓肠肌，阳性反应为踇趾背屈，有时伴其他足趾呈扇形展开，提示锥体束受损。由法国人戈登（Alfred Gordon）命名。
见载：神经病学名词
学科：神经病学_神经病学症状与体征
公布年度：2020

智能聚合

1.医学名词 第四分册
规范用词：戈登征
英语名：Gordon sign
学科：医学_神经病学
公布年度：1995

Somnambulism

ARTHROGRYPOSIS MULTIPLEX CONGENITA, DISTAL, TYPE IIA

UMLS CUI: C0220666

Semantic Types: Disease or Syndrome

Definitions (3)

Distal arthrogryposis type 3 is an autosomal dominant non-progressive myopathy with contractures of the hands, ankle and feet along with cleft palate and short stature. (SNOMEDCT_US)

Gordon syndrome, also known as distal arthrogryposis type 3, is an extremely rare multiple congenital malformation syndrome characterized by congenital contractures of hand and feet with variable degrees of severity of camptodactyly, clubfoot and, less frequently, cleft palate. Intelligence is normal but in some cases, additional abnormalities, such as short stature, kyphoscoliosis, ptosis, micrognathia, and cryptorchidism may also be present. Gordon syndrome, Marden-Walker syndrome and arthrogryposis with oculomotor limitation and electroretinal anomalies clinically and genetically overlap, and could represent variable expressions of the same condition. (ORPHANET)

Gordon征 – Gordon Syndrome – C0220666
戈登征 – Gordon Syndrome – C0220666
梦游症 – Sleep Walking – C0037672
……

Figure S1. Overview for mapping Chinese medical terms from real-world documents to the UMLS.

Step 1. Translate the Chinese medical term to English and calibrate the translation with the aid of medical dictionaries and web engines.

Step 2. Search the definition of Chinese medical terms as the preparation of mapping.

Step 3. Search candidate concepts and their definitions using *UMLS Metathesaurus Browser* according to calibrated translations.

Step 4. Compare term definitions between the Chinese medical term and its candidates and narrow down the candidates set to a short list.

[IMPORTANT] NOTE:

In some situations, it is hard for annotators to discern the mapping result by simply comparing string similarity between translations and candidate concepts because of the ambiguity of some UMLS concepts. Therefore, the definition of Chinese medical terms and candidate concepts play an important role to assist annotators to distinguish the similar candidate concepts.

For example, there are two highly similar candidate concepts for 肾囊肿 (Cyst of kidney). They are **Renal cyst** (C3887499) and **Cystic kidney** (C0022679). This is a hard case for selecting the mapping result without the aid of concepts definitions. According to the Metathesaurus of the UMLS, the definition of **Renal cyst** (C3887499) is that Renal cyst is a fluid filled sac in the kidney. And the definition of **Cystic kidney** (C0022679) is that Cystic kidney is a kidney containing one or more cysts. The definition of 肾囊肿 (Cyst of kidney) is same as the **Renal cyst** (C3887499). Thus, **Renal cyst** (C3887499) is selected as the mapping result.

Step 5. Screen the final candidates and select the mapping result following the mapping criteria (**Section 3**).

3. Mapping Criteria for Disambiguation

Although, most of Chinese medical terms can be mapped to the UMLS correctly with the aid of translations and definitions. There are still some challenges for annotators to choose the mapping result. The following criteria are principles to ensure the consensus for annotators when mapping terms to the UMLS.

Criterion 1. Concepts which are not tagged with T033|Finding semantic type have the priorities to be selected as the mapping results.

For example, **高眼压** (intraocular hypertension) which means a kind of ocular disorder with high intraocular pressure. Candidate concepts, **Ocular Hypertension** (C0028840) and **Raised intraocular pressure** (C0234708), are the candidate concepts as the mapping result for the **高眼压**. To distinguish **Ocular Hypertension** (C0028840) and **Raised intraocular pressure** (C0234708) for selecting the mapping result, semantic types of these two candidate concepts are considered. Semantic type of **Ocular Hypertension** (C0028840) is T047|Disease or Syndrome, and **Raised intraocular pressure** (C0234708)’s semantic type is

T033|Finding. Thus, **Ocular Hypertension** (C0028840) with T047|Disease or Syndrome is the final mapping result for the 高眼压 (intraocular hypertension) according to the criterion 2.

Criterion 2. Concepts which are included by SNOMED-CT vocabulary have the priorities to be selected as the mapping results.

For example, Concepts, **Malaria, Cerebral** (C0024534) and **malaria; cerebral** (C1403485), are candidate concepts for 脑型疟疾 (cerebral malaria). Semantic types of these two candidate concepts are both defined as T047|Disease or Syndrome. However, **Malaria, Cerebral** (C0024534) has been included in SNOMED-CT since 2002. **malaria; cerebral** (C1403485) is only included in ICPC2ICD10 vocabulary. Thus, under criterion 3, **Malaria, Cerebral** (C0024534) should be the final mapping result for the 脑型疟疾 (cerebral malaria).

Criterion 3. Medical terms that cannot be mapped to the UMLS are excluded.

For example, 地方性饮水型砷中毒 (endemic arsenism by water drinking) is a special endemic disease in China. This medical concept has not been included in any vocabularies of the UMLS. Thus, this kind of medical concepts should be excluded from the dataset.

Reference

[1] L. Deng, X. Zhang, T. Yang, M. Liu, L. Chen, and T. Jiang, “PIAT: An Evolutionarily Intelligent System for Deep Phenotyping of Chinese Electronic Health Records,” *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4142–4152, Aug. 2022, doi: 10.1109/JBHI.2022.3177421.

[2] S. Li *et al.*, “Deep Phenotyping of Chinese Electronic Health Records by Recognizing Linguistic Patterns of Phenotypic Narratives With a Sequence Motif Discovery Tool: Algorithm Development and Validation,” *J. Med. Internet Res.*, vol. 24, no. 6, p. e37213, Jun. 2022, doi: 10.2196/37213.