COVID Term: A Bilingual Terminology for COVID-19

Hetong Ma  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Liu Shen  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Haixia Sun  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Zidu Xu  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Li Hou  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Sizhu Wu  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

An Fang  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Jiao Li  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Qing Qian  
Institute of Medical Information/Library, Chinese Academy of Medical Sciences/Peking Union Medical College

Qian.qing@imicams.ac.cn  
https://orcid.org/0000-0002-9072-586X

Database

Keywords: COVID-19, COVID Term, COVID Terminology, bilingual, medical terminology

DOI: https://doi.org/10.21203/rs.3.rs-30923/v1
Abstract

Background The coronavirus disease (COVID-19), a pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has shown its destructiveness with more than one million confirmed cases and dozens of thousands of deaths, which is highly contagious and still spreading globally. World-wide studies have been conducted aiming to understand COVID-19 mechanism, transmission, clinical features, etc. A cross-language terminology of COVID-19 is essential for improving knowledge sharing and scientific discovery dissemination.

Methods We developed a bilingual terminology of COVID-19 with mapping Chinese and English terms. The terminology construction follows the workflow (1) Classification schema design; (2) Concepts and sub-concepts assignment; (3) Terminology editing strategy; (4) Terminology property development; (5) Online deployment. We built open access for the terminology named as COVID Term, providing search, browse, and download services.

Results The proposed COVID Term include 10 categories: disease, anatomic site, clinical manifestation, demographic and socioeconomic characteristics, living organism, qualifiers, psychological assistance, medical equipment, instruments and materials, epidemic prevention and control, diagnosis and treatment technique respectively. In total, COVID Terms covered 464 concepts with 724 Chinese terms and 887 English terms. All terms are openly accessible online (COVID Term: http://covidterm.imicams.ac.cn).

Conclusions COVID Term is a bilingual terminology focused on COVID-19, the epidemic pneumonia with a high risk of infection around the world. It will provide updated bilingual terms of the disease to help health providers and medical professionals retrieve and exchange information and knowledge in multiple languages. COVID Term was released in machine-readable formats (e.g., XML and JSON), which would contribute to the machine translation and advanced intelligent techniques.

Background

The coronavirus disease (COVID-19), a type of pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was known in December 2019[1]. With extremely high infectiousness, COVID-19 confirmed cases have been up to 83,597 in China till April 13, 2020[2]. During the period China has taken control measures such as increasing diagnostic testing rates, suspected cases isolation, investigating contacts of confirmed cases, clinical records and management, and restrictions on mobility[3], meanwhile requesting folks not to gather together, and must wear a mask, the interventions were effective. COVID-19 experienced an outbreak worldwide, till April 13, 2020, the confirmed cases in the whole world have been 1,773,084, together with 111,652 deaths[2].

Medical studies have been conducted to investigate COVID-19 globally. Clinical findings were analyzed via different test methodology[4], and in diverse groups, e.g. children, pregnant patients, regular patients, critically ill patients[5-9]. Modelling study were applied in different areas[10-13] and towards disparate directions, e.g. control strategy interventions, dynamic transmission, etc.[3, 12-14]. Clinical case study
were conducted in a specific region, or a specific domain e.g. risk factors\cite{15, 16}. Virus-related features and distribution analysis\cite{17} were also investigated. Liu et.al. completed an ontology towards coronavirus named Coronavirus Infectious Disease Ontology (CIDO) and made related analysis\cite{18}.

While more structured machine-readable data should be fed to the intelligent techniques such as deep learning model to realize further data mining and ontology construction, cross-lingual terminology is necessary to be created. Knowledge dissemination could be promoted through a cross-lingual professional system. Also, parallel machine-readable corpus would contribute to more applications, e.g. machine translation. In this study, we constructed a bilingual COVID terminology aiming to accelerate the COVID knowledge spreading across different countries and provide bilingual machine-readable data for intelligent techniques.

**Methods**

Considering the terminology data could be reused for ontology and knowledge graph construction, we referred to several ontology classification or workflow strategy \cite{19, 20}, and authoritative terminology system UMLS, SNOMED CT for terminology criteria establishment \cite{21, 22}. The designed workflow included five steps, respectively (1) Classification schema design (2) Concepts and the sub-concepts assignment (3) Terminology editing strategy (4) Terminology property development (5) Online deployment

**Classification schema design**

According to the research of mass COVID-19 data, together with suggestions from two experts in medical informatics, we developed 10 classification schema for the first level top nodes involving disease, anatomic site, clinical manifestation, demographic and socioeconomic characteristics, living organism, qualifiers, psychological assistance, medical equipment, instruments and materials, epidemic prevention and control, diagnosis and treatment technique.

**Concepts and the sub-concepts assignment**

COVID Term was designed as a 6-level structure, few terms were included in the sixth level, which was easier for users to search and browse. Too many layers would mislead and confuse users. Figure 1 demonstrated the hierarchical structure of COVID Term. The second layer contained body tissue structure, symptoms and signs, pathogen, route of transmission, clinical typing, extreme behaviors, medical protective items, specimen types, monitoring index, clinical trial, laboratory diagnostic method, imaging diagnostic method, etc. The third layer included protective bodywear, face shield, protective glasses, gown, coverall, mask, gloves, respiratory tract specimens, other specimens, etc. The fourth layer involved serum specimens, surgical mask, etc. The fifth layer covered disposable surgical mask, convalescent serum sample, and acute serum sample, etc. The sixth layer contained few virus types. Agile model was adopted during the data processing, i.e. adjust the structure by adding, altering or deleting specific substructures when necessary.
Terminology editing strategy

A bilingual terminology system towards a worldwide emergency disease was supposed to be correct, authoritative and highly correlated, where exact bilingual concept, semantic types, etc. should be demonstrated. Therefore, the resources we have taken use of were limited to authority publishment (website, report, document, etc.) e.g. the situation report or document from World Health Organization, journal articles (preprint, open access, etc.), nationwide regulation, policy document, professional books, etc. Bilingual terms were mostly extracted from bilingual WHO documents, textbooks, and related papers. Definitions were located from textbooks and related papers under most conditions. The data processing was performed on TBench, a work platform for cross-lingual terminology operation[23]. After data processing in each round, two examiners with professional background and related practice experience were invited to validate the accuracy of the terminology. A third party with clinical experts would be involved when disagreement was reached.

Terminology property development

Each term was assigned several properties including concept ID, semantic type, Chinese preferred term, English preferred term as the obligatory items, and Chinese synonym, English synonym, bilingual definition, definition source as alternative items. The processing dates and time would be automatically generated through the system. Among these properties, concept ID could be directly linked to other systems through automatic mapping, each definition should come with a source for users to look up to. Synonyms were not a prerequisite element but more synonyms would help with the search scope and term location.

Online deployment

Currently, the terminology has been updated twice, each with more term branches of abundant information. The up to date COVID resources e.g. the lancet coronavirus theme, NIH 2019 novel coronavirus theme, WHO COVID-19 theme[24-26], etc. were constantly followed by COVID team to provide most recent terminology. The terminology towards COVID-19 was named as COVID Term. We built a website for COVID Term, making it available for users to access. Earlier versions were also released on the PHDA(Population Health Data Archive)[27].

Results

COVID Term Overview

We constructed a 6-level bilingual terminology system for COVID-19 involving 464 concepts, 724 Chinese terms, 887 English terms, 464 Chinese preferred terms, 464 English preferred terms, 260 Non-preferred Chinese terms, 423 Non-preferred English terms, 42 Chinese definitions, and 5 English definitions. The first-level category was designed as disease, living organism, clinical manifestation, diagnosis and treatment technique, qualifiers, demographic and socioeconomic characteristics, epidemic prevention and
control, medical equipment, instruments and materials, psychological assistance, and anatomic site. Partial hierarchical structure and corresponding terms were shown in Figure 2.

**Concept Distribution**

In this 6-level structure, all first-level terms were root nodes with leafs. There were 10 first-level terms, same as first-level categories, 104 second-level terms, 180 third-level terms, 138 fourth-level terms, 31 fifth-level terms, and 9 sixth-level terms. Sixth-level leaf nodes only existed in category ‘Disease’, ‘Living Organism’, and ‘Anatomic Site’. One example was ‘Living Organism – pathogen – virus – coronavirus – human coronaviruses – severe acute respiratory syndrome coronavirus 2’. Concept distribution in different categories was calculated. Concepts in ‘Clinical Manifestation’, ‘Diagnosis and Treatment Technique’, ‘Epidemic Prevention and Control’ ranked top 3, which together accounted for over half concepts, with proportion respectively 20%, 18%, 14%. Concept distribution was shown in Figure 3.

**Terminology Property Analysis**

We performed data statistics on each property of every category including concept, bilingual term count, bilingual synonym count, and bilingual definition count in descending order of concept. Concepts in category ‘Diagnosis and Treatment Technique’ ranked first with 94 terms, together with 99 Chinese synonyms and 125 English synonyms.

Most English terms took a major place in all properties among different categories. Few English definitions were included in the system. Category ‘Living Organism’ showed the most definitions as 15. Categories with Chinese synonyms over 50 were ‘Diagnosis and Treatment Technique’ and ‘Clinical Manifestation’, as shown in Figure 4.

**Semantic Type Distribution**

We assigned a semantic type for most of the concepts, which generally was the parent node or the grandfather node of that concept. For example, the semantic type of either latex gloves or gloves was labelled ‘Medical Protective Items’, representing its semantic meaning. There were 18 semantic types in total as illustrated. As seen in Figure 5, the top 3 semantic types were ‘Epidemic Prevention and Control’, ‘Disease’, and ‘Medical Protective Items’. Concepts in top half semantic types covered nearly 80% of the whole terms. There was only one terminology in pathological examination i.e. biopsy. Different clinical stages were classified into the semantic type ‘Clinical Trial’.

**Website Demonstration**

We built a website for users to access COVID Term, which provided browsing, searching, and downloading, etc. services. On the homepage, users could choose to search, navigate (browse), download, and leave feedback online. Users could also see the visit number statistics for different services. Concept, term, and relation numbers were shown on the same page. The website also provided
In the search function, users could search wanted terms through general or advanced search, where search in different language terms, definitions, exact matches between term and keyword, etc. On the navigation page, users could select each level concept and their subconcepts, details were provided on the right with concept ID, preferred terms, non-preferred terms, definitions, hierarchical relationship, and semantic types. A simple knowledge graph was also shown in the visualization branch, illustrating the relationship of each concept and related terms, as shown in Figure 7. Users could download the whole dataset in Excel, XML, JSON format.

Discussion

Application

COVID-19, an extremely destructive disease confirmed for several months, has caused over 100 thousand deaths, and more than 1.5 million confirmed cases[2]. This emergency leads to a common scenario that many patients who were lack of medical resources had to suffer from the disease. With the constant enlargement of the epidemic coverage, related data and information in various types have grown exponentially.

We constructed a 6-level 10-category COVID-focused terminology system named COVID Term by collecting and integrating highly related bilingual concepts, synonyms, definitions, and semantic types. 464 bilingual preferred terms, 724 Chinese terms, 887 English terms were included in the terminology system, where each term was classified into a category involving Diagnosis and Treatment Technique, Epidemic Prevention and Control, Psychological Assistance, Clinical Manifestation, Qualifiers, Disease, Anatomic Site, Living Organism, Medical Equipment, Instruments and Materials, and Demographic and Socioeconomic Characteristics. In addition, we provided simple knowledge graphs to illustrate their relationship and built a website so that terminology dataset could be open accessible to all users.

COVID Term is a terminology dataset applicable for people from all over the world, despite careers, ages, and nationalities. Front-line physicians, clinicians, and other clinical staff could look up to the Term as a dictionary, quickly obtaining appropriate ways of specific professional vocabulary aiming at COVID-19 and its definition. Medical informatics scientists, analysts, and other data processing technicians could take maximum advantage of COVID Term via definition extraction, terminology matches, ontology construction, and related machine translation. The bilingual terminology would promote knowledge dissemination and information exchange in COVID-19.

Limitations and future studies

Currently, there are only 42 Chinese definitions and 5 English definitions included in the system. More definitions need to be integrated. With the growing data generated from all over the world, more
classifications might be included and more relationships might need to be designed. In the future, we would continue collecting related terminology, constructing more mature system, and designing more interfaces applicable for various platforms.

Conclusions

To promote the COVID-19 related information spreading, data reuse through intelligent techniques, to realize knowledge graph construction, and to synchronize Chinese up-to-date outcome in epidemic prevention, we constructed a COVID focused terminology system at vocabulary level, named as COVID Term. COVID Term is a bilingual terminology system including 464 bilingual COVID related terminology, each with bilingual preferred term, concept ID, and other alternative properties such as bilingual synonyms, bilingual definitions, and semantic type. All terms were accessible online through the website, which provided search, navigate, browse, download, and visualization, etc. services. COVID Term would promote knowledge dissemination and contribute to advanced intelligent techniques.

Abbreviations

**SNOMED CT**: Systematized Nomenclature of Medicine–Clinical Terms; **UMLS**: Unified Medical Language System; **NIH**: National Institutes of Health; **XML**: Extensible Markup Language; **JSON**: JavaScript Object Notation;

Declarations

Authors’ contributions

All the authors participated in the study design and manuscript writing. QQ conducted the study. All the authors participated in designing the terminology construction schema. HM, LS, HS completed the terminology editing and result analysis. ZX, LH, SW, AF, JL participated in the terminology validation. All the authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The dataset generated and analyzed during the current study is publically available via [http://covidterm.imicams.ac.cn](http://covidterm.imicams.ac.cn).
Competing interests

The authors declare that they have no competing interests.

Funding

This research is funded by the National Key Research and Development Program of China (Grant No. 2016YFC0901901), the Chinese Academy of Medical Sciences (Grant No. 2017-I2M-3-014), and the National Population Health Scientific Data Center.

Acknowledgements

Not applicable.

References

1. SARS-CoV-2.WHO. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed 13 April 2020.
2. WHO: Coronavirus disease 2019 (COVID-19) Situation Report – 84. 2020.
3. Kraemer MUG, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, Open C-DWG, du Plessis L, Faria NR, Li R et al: The effect of human mobility and control measures on the COVID-19 epidemic in China. Science (New York, NY) 2020.
4. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, Fan Y, Zheng C: Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. The Lancet Infectious Diseases 2020, 20(4):425-434.
5. Qiu H, Wu J, Hong L, Luo Y, Song Q, Chen D: Clinical and epidemiological features of 36 children with coronavirus disease 2019 (COVID-19) in Zhejiang, China: an observational cohort study. The Lancet Infectious Diseases.
6. Chen H, Guo J, Wang C, Luo F, Yu X, Zhang W, Li J, Zhao D, Xu D, Gong Q et al: Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: a retrospective review of medical records. The Lancet 2020, 395(10226):809-815.
7. Yu N, Li W, Kang Q, Xiong Z, Wang S, Lin X, Liu Y, Xiao J, Liu H, Deng D et al: Clinical features and obstetric and neonatal outcomes of pregnant patients with COVID-19 in Wuhan, China: a retrospective, single-centre, descriptive study. The Lancet Infectious Diseases.
8. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y et al: Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet 2020, 395(10223):507-513.
9. Yang X, Yu Y, Xu J, Shu H, Xia Ja, Liu H, Wu Y, Zhang L, Yu Z, Fang M et al: Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered,
retrospective, observational study. *The Lancet Respiratory Medicine.*

10. Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boëlle P-Y, D’Ortenzio E, Yazdanpanah Y, Eholie SP, Altmann M et al: Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *The Lancet* 2020, **395**(10227):871-877.

11. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, Flashe S, Clifford S, Pearson CAB, Munday JD et al: The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health.*

12. Koo JR, Cook AR, Park M, Sun Y, Sun H, Lim JT, Tam C, Dickens BL: Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *The Lancet Infectious Diseases.*

13. Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD et al: Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases.*

14. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, Munday JD, Kucharski AJ, Edmunds WJ, Sun F et al: Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health* 2020, **8**(4):e488-e496.

15. Ghinai I, McPherson TD, Hunter JC, Kirking HL, Christiansen D, Joshi K, Rubin R, Morales-Estrada S, Black SR, Pacilli M et al: First known person-to-person transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the USA. *The Lancet.*

16. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X et al: Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet* 2020, **395**(10229):1054-1062.

17. To KK-W, Tsang OT-Y, Leung W-S, Tam AR, Wu T-C, Lung DC, Yip CC-Y, Cai J-P, Chan JM-C, Chik TS-H et al: Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. *The Lancet Infectious Diseases.*

18. Liu Y, Chan W, Wang Z, Hur J, Xie J, Yu H, He Y: Ontological and Bioinformatic Analysis of Anti-Coronavirus Drugs and Their Implication for Drug Repurposing against COVID-19. *Preprints 2020* 2020, 2020030413.

19. Kim HH, Park YR, Lee KH, Song YS, Kim JH: Clinical MetaData ontology: a simple classification scheme for data elements of clinical data based on semantics. *BMC Medical Informatics and Decision Making* 2019, **19**(1):166.

20. Hier DB, Brint SU: A Neuro-ontology for the neurological examination. *BMC Medical Informatics and Decision Making* 2020, **20**(1):47.

21. Bodenreider O: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004, **32**(Database issue):D267-270.

22. SNOMED Clinical Terms. [https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html](https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html). Accessed 13 April 2020.
23. Deng P, Ji Y, Shen L, Li J, Ren H, Qian Q, Sun H: TBench: A Collaborative Work Platform for Multilingual Terminology Editing and Development. Stud Health Technol Inform 2019, 264:1449-1450.

24. The lancet coronavirus theme. https://www.thelancet.com/coronavirus. Accessed 13 April 2020.

25. NIH 2019 novel coronavirus theme [https://www.ncbi.nlm.nih.gov/research/coronavirus/]. Accessed 13 April 2020.

26. WHO COVID-19 theme. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed 13 April 2020.

27. Population Health Data Archive. http://www.ncmi.cn/index.html. Accessed 13 April 2020.

Figures
Disease [疾病]
- COVID-19 [新型冠状病毒肺炎]
- Respiratory System Disorder [呼吸系统疾病]
  - Pneumonia [肺炎]
  - Viral Pneumonia [病毒性肺炎]
    - Infectious Mononucleosis Pneumonia [传染性单核细胞增多性肺炎]
    - Mononuclear Interstitial Pneumonia [单核间质性肺炎]
- Comorbidities and Complications [合并症和并发症]

Living Organism [生物体]
- Pathogen [病原体]
  - Virus [病毒]
    - Coronavirus [冠状病毒]
      - Human Coronaviruses [人类冠状病毒]
        - Severe Acute Respiratory Syndrome Coronavirus 2 [新型冠状病毒]

Clinical Manifestation [临床表现]
- Imaging Examination Abnormal [影像学检查异常]
- Symptoms and Signs [症状体征]
- Test and Examination Abnormal [检验检查异常]

Diagnosis and Treatment Technique [诊疗技术]
- Clinical Trial [临床试验]
- Imaging Diagnostic Method [影像学诊断方法]
- Laboratory Diagnostic Method [实验室诊断方法]
- Monitoring Index [监测指标]
- Pathological Examination [病理检查]
- Specimen Types [标本类型]

Qualifiers [限定语]
- Clinical Typing [临床分型]
- Route of Transmission [传播途径]

Demographic and Socioeconomic Characteristics [人口学及社会经济学特征]
- Population [人群]

Epidemic Prevention and Control [疫情防控]

Medical Equipment, Instruments and Materials [医用设备、器械和材料]
- Medical Protective Items [医用防护物品]
  - Coverall [工作服]
  - Mask [口罩]
    - Surgical Mask [外科口罩]
      - Disposable Surgical Mask [一次性医用外科口罩]

Psychological Assistance [心理救助]
Figure 1

Hierarchical structure of COVID Term
Figure 2

COVID Term structure illustration
Figure 3
Concept Distribution of COVID Term

Figure 4
Term Property Distribution of COVID Term
Figure 5

Semantic type distribution of COVID Term
Figure 6

COVID Term website homepage

| Detailed Information       | Visualization |
|----------------------------|---------------|
| Concept ID: 282092188144234501 |               |
| CCID: CC00001547            |               |
| Preferred Terms: Aerosol Transmission |       |
| Non-preferred Terms: Aerosol Transmission Route |   |
| Airborne Spread During Aerosol-generating Medical Procedures |     |
| Airborne Through Aerosols |               |
| Definitions: Aerosol Transmission |         |
| Hierarchical Relationship: Qualifiers | Route of Transmission |
| Semantic Types: Route of Transmission / 传播途径 |   |

Figure 7

Concept details in COVID Term