Letter to the Editor

Authors' Reply: Baseline Fidelity and Evaluation Independence in Symptom Terminology Normalization

Junyu Yao, MCM; Dong Cao, PhD

School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou, Guangdong, China

Corresponding Author:

Dong Cao, PhD
School of Medical Information Engineering
Guangzhou University of Chinese Medicine
232 Outer Ring East Road, Panyu District
Guangzhou, Guangdong 510006
China
Phone: 86 13246852860
Email: caodong9@163.com

Related Articles:
Comment in: https://medinform.jmir.org/2026/1/e113829
Comment on: https://medinform.jmir.org/2026/1/e85825

JMIR Med Inform 2026;14:e115495; doi: 10.2196/115495

Keywords: symptom terminology; traditional Chinese medicine; natural language processing; named entity recognition; named entity normalization; pretrained language model

To the Editor

We thank the reader [1] for the careful assessment of our study [2] and the constructive comments, which provide an opportunity to explain our methodological choices.

Our description of PLM-*ICD* (pretrained language model framework for *International Classification of Diseases* coding) as producing a single output referred to the implementation adapted for our normalization task. PLM-*ICD* was originally developed for document-level *International Classification of Diseases* coding and supports multilabel prediction [3]. As it was not specifically designed for multi-implication terminology normalization, we adapted its inference rule while retaining multilabel training. Our implementation returned at most 1 standard-term label, or no label for unmatched terms, following a single-best-match approach consistent with Chen et al [4]. Sequence generation and the multitask candidate generator (MTCG) were designed to normalize compound terms containing multiple concepts. We retained their multiple-term outputs: sequence generation produced several standard terms, while our adapted MTCG implementation combined candidate retrieval with term-count prediction.

PLM-*ICD* was trained only on the texts and corresponding standard-term labels from the 400 training records. Its multilabel classifier had no separate candidate-term matching or reranking module and therefore required no additional positive or negative term pairs. The BERT (bidirectional encoder representations from transformers) candidate-reranking module in sequence generation and the candidate semantic-matching module in our adapted MTCG implementation learn whether raw terms match candidate standard terms. Both modules were trained on the same 3350 positive and negative term pairs as the BERT semantic-alignment module in our framework.

We augmented terms from all 500 records before dividing them into 400 training and 100 test records. Of the 3350 alignment examples, 205 were derived from the 100 test records, either directly from the original texts or through data augmentation. The evaluation reported in Table 6 of our paper [2] included both core symptom terms and supplementary semantic attributes. In the single-implication test set, 29 fine-grained raw terms contained in the test instances also appeared in the alignment training data in their original or augmented form. The corresponding counts were 41 for the multi-implication test set and 128 for the mixed test set.

Sallauka et al [5] also studied symptom extraction using medical named entity recognition in low-resource languages, where manually annotated data are scarce. Given the similar research focus and constraints on annotated resources, we followed their reported sequence of augmentation before data splitting. This was intended to increase expression coverage under limited annotation resources. The 3350 alignment

examples, both positive and negative, consisted of raw term–standard term pairs. Raw terms came from the original text of all 500 records or manually augmented texts. The augmented texts were prepared manually using 5 predefined rules (replace, merge, insert, combine, and copy). Standard terms were drawn from authoritative international classifications and terminology systems, medical texts, the most commonly used expressions in the records, or formulations developed through expert discussion.

We selected the bidirectional long short-term memory–conditional random field model based on the test results shown in Table 5 [2], considering its comparable performance in recognizing structural entities, model complexity, and computational cost. It served as the entity-recognition component, while a BERT binary classifier performed alignment. The study focused on demonstrating the effectiveness of split-then-concatenate normalization, rather than identifying the base model with the highest test scores.

We agree with the reader's recommendations for strengthening the evaluation design. These overlaps and the use of test results for model selection limit the independence of the current evaluation. Once sufficient additional annotated medical records are available, we plan to use patient-level training, validation, and test splits, drawing corpus-derived resources and augmented examples only from the training data. Validation data would guide model selection before the finalized pipeline is evaluated on independent test records. By incorporating component ablations and separate analyses of expressions seen and unseen during training, we aim to clarify component contributions and generalization to new expressions.

Generative AI Use

No generative AI tools were used in the writing or editing of this manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (grant 2023YFC3503002), the 2026 "Unveiling and Commanding" Project of the School of Medical Information Engineering at Guangzhou University of Chinese Medicine (grant 2026-7), and the Platform Project of the Big Data Research Center for Traditional Chinese Medicine at Guangzhou University of Chinese Medicine (grant A1-2601-25-439-127Z102).

Data Availability

The datasets generated and analyzed during this study are not publicly available due to institutional data governance policies and patient confidentiality requirements. Deidentified data supporting the findings of this study may be available from the corresponding author upon reasonable request, subject to approval by the relevant institutional authorities and compliance with local data protection regulations.

Authors' Contributions

JY contributed to conceptualization, data curation, formal analysis, methodology, software, validation, writing (original draft), and writing (review and editing). DC contributed to conceptualization, funding acquisition, methodology, project administration, resources, supervision, and writing (review and editing).

Conflicts of Interest

None declared.

References

1. He Q. Baseline fidelity and evaluation independence in symptom terminology normalization. JMIR Med Inform. 2026;14:e113829. [doi: 10.2196/113829]
2. Yao J, Gou X, Lai W, et al. Symptom terminology normalization in traditional Chinese medicine: development and evaluation of a 2-stage deep learning framework based on fine-grained semantic classification. JMIR Med Inform. Sep 25, 2026;14:e85825. [doi: 10.2196/85825] [Medline: 42789749]
3. Huang CW, Tsai SC, Chen YN. PLM-ICD: automatic ICD coding with pretrained language models. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. 2022:10-20. [doi: 10.18653/v1/2022.clinicalnlp-1.2]
4. Chen L, Varoquaux G, Suchanek FM. A lightweight neural model for biomedical entity linking. AAAI. 2021;35(14):12657-12665. [doi: 10.1609/aaai.v35i14.17499]
5. Sallauka R, Arioz U, Rojc M, Mlakar I. Weakly-supervised multilingual medical NER for symptom extraction for low-resource languages. Applied Sciences. 2025;15(10):5585. [doi: 10.3390/app15105585]

Abbreviations

BERT: bidirectional encoder representations from transformers
MTCG: multitask candidate generator
PLM-ICD: pretrained language model framework for International Classification of Diseases coding

Edited by Andrew Coristine; This is a non–peer-reviewed article; submitted 28.Oct.2026; initial comments to author 06.Nov.2026; final revised version received 12.Nov.2026; accepted 12.Nov.2026; published 26.Nov.2026

Please cite as:
Yao J, Cao D
Authors' Reply: Baseline Fidelity and Evaluation Independence in Symptom Terminology Normalization
JMIR Med Inform 2026;14:e115495
URL: https://medinform.jmir.org/2026/1/e115495
doi: 10.2196/115495

© Junyu Yao, Dong Cao. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.