JPPB: Automatic Construction of a Soft-Labeled Japanese Patient Phrase Bank for Symptom Normalization

Tomohiro Nishiyama, Mana Kuramoto, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology
Nara, Japan
{nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp
kuramoto.mana.kj4@naist.ac.jp

Abstract

Patient-generated symptom expressions are linguistically diverse, often deviating from standardized medical terminology. This paper introduces the Japanese Patient Phrase Bank (JPPB), the first automatically constructed phrase-level normalization resource for Japanese patient language. JPPB introduces an embedding-based soft labeling framework that transforms traditional one-to-one dictionary mappings into graded and ambiguity-aware associations. This framework represents a shift from word-level to phrase-level normalization in Japanese. The resource covers 7,035 phrase‑term pairs across 412 symptoms. Evaluation on the KEEPHA and MedNLP-SC datasets shows that soft labels consistently improve Top-1 accuracy and better approximate gold label distributions compared with hard labels. While LLM-based normalization achieved the highest scores, JPPB provides a lightweight and transparent alternative suitable for local deployment. This work demonstrates that large-scale, automatically generated phrase banks can achieve competitive performance relative to manually curated resources and serve as practical, scalable resources for medical natural language processing in Japanese.

Keywords: patient expression normalization, medical NLP, resource construction, Japanese language, soft labeling

1. Introduction

In clinical settings, patients rarely describe their symptoms using standardized medical terminology (Zeng and Tse, 2006). Instead, they often rely on colloquial, metaphorical, or ambiguous expressions (e.g., describing “fever” as “my body feels hot”), making it difficult to directly map their narratives to canonical medical terms. Such variation reflects natural differences in age, education, and sociolinguistic background, but poses fundamental challenges for medical natural language processing (NLP) tasks. Inaccurate normalization of patient expressions can lead to symptom misclassification, incomplete detection of adverse events, and biased analyses in observational studies, ultimately undermining the reliability of medical AI systems (Zeng Qing et al., 2001; Hayes et al., 2017).

Over the past decade, computational approaches to patient-generated text have evolved from rule-based and lexicon-driven systems (Zeng Qing et al., 2001; Zeng et al., 2002) to embedding-based and LLM-based models (Ibrahim et al., 2021; Yao et al., 2024), substantially improving the automatic understanding and normalization of lay‑medical language. However, despite these advances, real-world deployment of such models remains limited by privacy, computational, and governance constraints (Dennstädt et al., 2025; Sarker et al., 2024), highlighting the continuing need for lightweight and interpretable lexical resources that can operate locally without reliance on external APIs. While many studies on medical NLP have concentrated on English, systematically constructed resources remain scarce for other languages, hindering the broader applicability of existing methods. While probabilistic lexicons and word-sense distribution models have been explored in general-domain English text (Erk et al., 2009), they rarely address patient-generated medical language in non-English settings. In contrast, JPPB targets this gap by constructing a soft-label dictionary for Japanese patient expressions and systematically quantifying what we term clinical ambiguity. Here, clinical ambiguity refers to cases in which a patient expression can plausibly correspond to multiple canonical symptom terms. Japanese, in particular, exemplifies these challenges: its patient language frequently employs onomatopoeia (e.g., mukamuka for nausea) and polysemous adjectives (e.g., shindoi for tired or sluggish), whose meanings depend strongly on context, making extraction and normalization highly non-trivial (Nishiyama et al., 2024). Although several Japanese patient lexicons have been developed, they are often limited in coverage, outdated, or inconsistent in handling colloquial or emerging expressions, such as newly reported drug side effects or evolving public-health concerns, thereby constraining the generalizability and reliability of medical NLP systems in Japanese healthcare.

To address this challenge, we present the Japanese Patient Phrase Bank (JPPB), a sys-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7816–7828
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

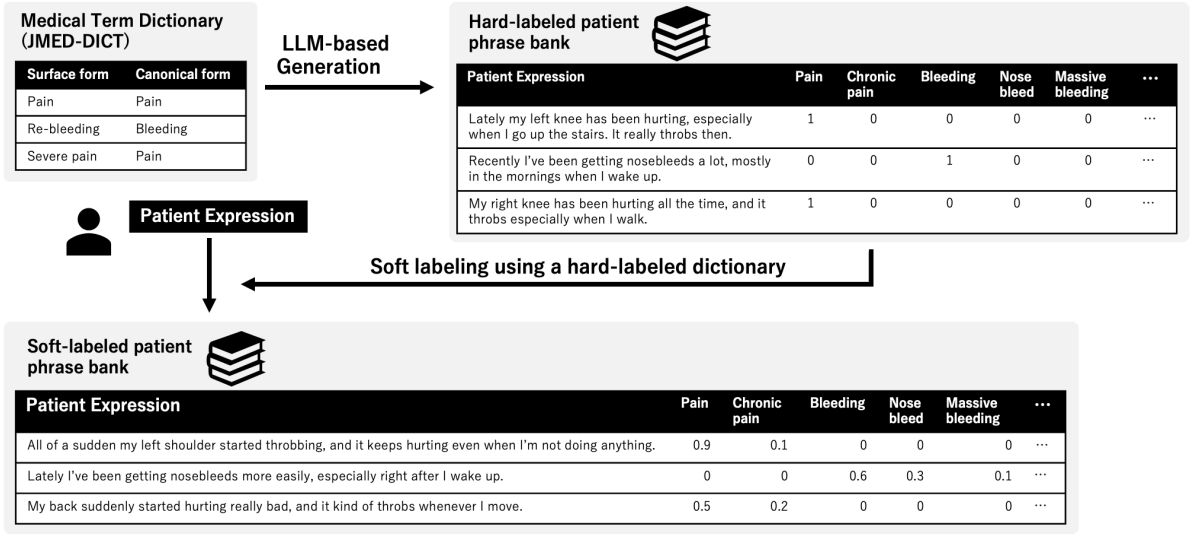

Figure 1: The construction of JPPB. Canonical symptom terms from the medical term dictionary served as prompts for LLM-based generation, which produced diverse patient expressions directly associated with their source canonical terms, thereby forming a hard-labeled patient phrase bank. Soft labeling was then applied to create the soft-labeled patient phrase bank, which associates each expression with multiple canonical terms scored by the frequency of similar expressions.

tematically developed resource that maps diverse patient symptom expressions to canonical medical terms. As shown in Fig. 1, JPPB integrates two complementary processes: (1) hard labeling, which generates multiple patient expression variants for each canonical term, and (2) soft labeling, which automatically propagates semantically related standardized terms via embedding-based similarity.

This framework enables automatic construction of a phrase-level normalization resource that systematically captures lexical diversity and linguistic ambiguity in Japanese patient language. Unlike previous resources such as the Large-Scale Patient Expression Dictionary[1] and other manually curated patient lexicons, which are typically limited to one-to-one, word-level mappings, JPPB introduces soft labels that capture graded semantic relationships between expressions and canonical terms, providing a scalable and ambiguity-aware alternative to existing lexicon-based resources.[2]

2. Related Work

2.1. Patient Expression Resources in Other Languages

Existing resources for patient-generated expressions have been primarily developed for English. Early initiatives such as the *Consumer Health Vocabulary (CHV)* (Zeng and Tse, 2006) provided mappings between lay and professional medical terms, mainly at the word level. However, subsequent analyses revealed that many consumer health expressions do not map cleanly to existing medical ontologies (e.g., UMLS) and often represent uniquely lay concepts grounded in everyday health perceptions (Keselman et al., 2008). To address such lexical gaps, embedding-based approaches have been proposed to automatically expand consumer vocabularies from online health forums (Gu et al., 2019). Nevertheless, they still operate mainly at the word level and do not capture contextual variation in patient expressions.

Other corpora such as *CADEC* (Karimi et al., 2015) and *PsyTAR* (Zolnoori et al., 2019) extended the scope to sentence-level patient narratives; their annotations primarily target within-sentence ADE mentions, i.e., they are designed for entity or relation modeling rather than for constructing a reusable phrase-level normalization lexicon. More recent studies have moved toward data-centric approaches that leverage large language models (LLMs) to bridge lay and professional language, such as the *README* framework (Yao et al., 2024).

However, even in English, most normalization datasets remain lexically focused, providing mappings at the word level rather than capturing the meaning of entire phrases. Comprehensive phrase-level normalization resources are still scarce, and existing datasets are not directly reusable for languages with distinct linguistic char-

[1] https://sociocom.naist.jp/patient-dic/

[2] The resource is freely available under a CC-BY license at https://github.com/Tomohiro-git/JPPB.

acteristics.

2.2. Patient Expression Resources in Japanese

In contrast, resources for Japanese patient language remain limited in both coverage and granularity. Lexical databases such as *J-MeDic* (Ito et al., 2018) and the *Patient Expression Dictionary* (Nishidani et al., 2021) provide mappings between surface forms and canonical terms. While these resources serve as important foundations for clinical NLP, they cannot adequately represent compositional symptom descriptions or context-dependent variants common in patient speech.

Recent work has explored patient-oriented analyses within broader clinical NLP settings (Ohno et al., 2024; Nishiyama et al., 2024), yet there is still no systematically constructed phrase-level resource for Japanese. Furthermore, existing dictionaries rely on manual updates, which restrict scalability and delay adaptation to emergent expressions seen in user-generated data.

The *Japanese Patient Phrase Bank (JPPB)* addresses this gap by constructing a large-scale phrase-level normalization dictionary that introduces a soft-labeling framework extending traditional one-to-one mappings. By leveraging embedding similarity during construction, JPPB extends traditional word-level dictionaries into a flexible, score-based framework that captures lexical diversity and linguistic ambiguity, offering the first systematically developed resource for Japanese patient language.

3. Construction of JPPB

This section outlines the methodology for constructing the JPPB.

3.1. Extraction of Canonical Symptom Terms

We began by extracting canonical symptom terms from JMED-DICT, a large-scale Japanese medical terminology dictionary that provides standardized mappings between surface forms and canonical medical terms[3]. Specifically, we selected symptom expressions corresponding to ICD-10 codes starting with "R," as this category is reserved for symptoms, signs, and abnormal clinical findings rather than disease names. Focusing on symptom-level expressions, we aimed to construct a resource that is more closely aligned with how patients actually describe their conditions. This filtering resulted in 783 pairs of surface forms (as actually used in clinical documentation) and their corresponding canonical forms.

[3] https://sip3-d2.naist.jp/jmed-dict.html

3.2. Generation of Patient Expressions

To capture the diversity of real-world patient language, we generated multiple variations of patient expressions for each canonical term using GPT-4.1-mini. We adopted a persona-based generation approach, creating three personas: "general," "child," and "elderly." For each persona, three patient expressions were automatically generated using a predefined structured prompt with an LLM (see Figures 3 and 4 for details), ensuring variation primarily in age and communication style.

3.3. Label Assignment and Finalization

We constructed two types of dictionaries:
Hard-label: Each generated patient expression is paired with a single canonical medical term.
Soft-label: For each generated phrase, we first retrieved the top 10 most similar patient expressions using embedding-based semantic similarity with a multilingual MiniLM model (Reimers and Gurevych, 2019). The canonical terms associated with these similar expressions were aggregated, and a label score was assigned according to their frequency. Each occurrence contributed 0.1 to the score, effectively normalizing the maximum score to 1.0 when a term appeared in all retrieved expressions. This allows each patient's expression to be associated with multiple potential canonical terms, weighted by their semantic proximity.

After filtering out duplicate entries, the final resource consisted of 7,035 unique pairs in the hard- and soft-label dictionaries.

4. Dataset Analysis

This section provides an analysis of the JPPB, corpus statistics, label distribution, and coverage in comparison with existing resources.

4.1. Corpus Statistics

The final JPPB is released as a soft-label dictionary, where each patient-generated expression is associated with a distribution over canonical symptom terms. This representation enables the resource to capture linguistic ambiguity and variation in patient language. For evaluation purposes, we also constructed a hard-label version consisting of 7,035 phrase-canonical pairs.

Corpus-level statistics are summarized in Table 2. Both dictionaries cover 412 canonical terms, but the soft-label version exhibits higher variability: on average, each canonical term is linked to over

Resource	Language	Unit	Construction	Label Type
CHV (Zeng and Tse, 2006)	English	Phrase	Manual	Hard
J-MeDic (Ito et al., 2018)	Japanese	Word	Manual	Hard
Patient Expression Dict. (Nishidani et al., 2021)	Japanese	Word	Manual	Hard
JPPB (ours)	Japanese	Phrase	Automatic	Soft

Table 1: Comparison between JPPB and existing resources.

96 surface variants, reaching up to 1,015 at maximum. By contrast, the hard-label version shows a smaller range of variants, with an average of 17 and a maximum of 261. These statistics demonstrate that the soft-label dictionary provides a more flexible mapping between patient expressions and canonical terms.

We quantified lexical ambiguity as the entropy of the soft-label probability distribution: $H(x) = -\sum_i p_i \log p_i$, where p_i denotes the probability assigned to the i-th canonical term, and log denotes the natural logarithm. Across all entries in JPPB, the entropy ranged from 0 to 2.303, with a mean of 1.483 and a median of 1.557. These values indicate that most expressions were associated with two to five canonical terms, reflecting moderate ambiguity. Representative examples of such ambiguous expressions are shown in Table 3.

To further illustrate the distribution of variants in the soft-label dictionary, we examined the canonical terms with the largest number of associated patient expressions, as shown in Table 4. Common symptoms such as pain and bleeding, as well as findings like tumor, exhibit particularly high numbers of variants.

4.2. Label Distribution

We evaluated how well the hard- and soft-label dictionaries approximate the gold label distributions in the test datasets described in Section 5.1. For each test instance, the gold distribution was defined as a uniform distribution over the set of annotated canonical terms. The predicted distribution was obtained from the top-1 retrieved dictionary entry, using its label-score vector normalized to sum to 1. To quantify similarity, we used cosine similarity, Jensen‑Shannon (JS) distance, and Kullback‑Leibler (KL) divergence.

Statistic	Hard	Soft
Avg. variants per canonical	17.1	96.4
Median variants per canonical	9	76
Min variants per canonical	9	9
Max variants per canonical	261	1015

Table 2: Comparison of corpus statistics between the hard- and soft-label dictionaries.

As shown in Table 5, the soft-label dictionary consistently achieves higher cosine similarity and lower divergence values than the hard-label dictionary across both datasets. This indicates that soft labels not only improve normalization accuracy but also better capture the inherent ambiguity of patient-generated expressions.

Additionally, we visualized the label distributions using t-SNE in Fig. 2. The hard-label representation forms distinct and clearly separated clusters, reflecting its binary assignments and strict one-to-one mappings. In contrast, the soft-label representation appears broader and more overlapping, indicating probabilistic associations that capture ambiguous boundaries between concepts.

4.3. Persona-level Analysis

To evaluate stylistic consistency across personas, we analyzed generated expressions from the *child*, *general*, and *elderly* settings. We measured the average sentence length and the politeness rate, defined as the proportion of sentences ending with predicate forms such as “です (*desu*)” or “ます (*masu*)”.

The median sentence length ranged from 40 to 42 characters, showing no substantial difference across personas (see Appendix Table 7). This indicates that all persona settings produced sentences of comparable length. Politeness rates exceeded 99% for all groups, indicating consistent use of formal spoken style regardless of persona. These results suggest that while the model maintains a polite and coherent tone, it does not yet reflect clear stylistic differentiation among personas.

Although the persona prompts (*child*, *general*, *elderly*) were expected to influence the diversification of expressions, the generated sentences showed no clear difference in length or politeness. This outcome likely reflects two factors: (1) the inherent limitation of current LLMs, which respond weakly to high-level persona conditioning without explicit stylistic constraints (Zheng et al., 2024), and (2) the appropriateness of polite speech in the modeled scenario, where users in Japanese clinical settings naturally address healthcare professionals with formality regardless of age or identity. Taken together, these findings suggest that the limited stylistic variation across personas re-

Patient expression (translated)	Soft-label distribution over canonical terms
おへその上あたりが重たい感じがして、何か食べると痛くなるんです。最近、特に気になることが増えました。*(I feel a heavy sensation above my navel, and it hurts when I eat. Recently, it has become more noticeable.)*	疼痛 (pain): 0.1, 心窩部痛 (epigastric pain): 0.1, 心窩部不快 (epigastric discomfort): 0.1, 腹部膨満 (abdominal distension): 0.1, 嚥下困難 (dysphagia): 0.1, 上腹部痛 (upper abdominal pain): 0.1, 腹部不快感 (abdominal discomfort): 0.1, 腫瘍 (tumor): 0.1, *etc.*
運動しているときに、急に足のすねのあたりがつって、動けなくなりました。ちょっとおかしいです。*(While exercising, my lower leg suddenly cramped, and I couldn't move. It felt strange.)*	痙攣 (cramp): 0.1, こむら返り (muscle cramp): 0.1, 筋力低下 (muscle weakness): 0.1, 四肢脱力 (limb weakness): 0.1, 歩行障害 (gait disturbance): 0.1, 歩行異常 (abnormal gait): 0.1, 不随意運動症 (involuntary movement): 0.1, 跛行 (limping): 0.1, *etc.*

Table 3: Examples of high-entropy expressions exhibiting genuine clinical ambiguity. Each case shows multiple plausible canonical terms reflecting overlapping interpretations of patient language.

Canonical term	Hard	Soft
疼痛 (Pain)	261	1015
腫瘤 (Tumor)	189	875
癌性疼痛 (Cancer pain)	63	503
出血 (Bleeding)	170	430
多臓器不全 (Multiple organ failure)	54	369
呼吸困難 (Dyspnea)	45	348
壊死 (Necrosis)	72	313
胸部異常陰影 (Chest abnormal shadow)	90	308
全身性炎症反応症候群 (Systemic inflammatory response syndrome)	45	299
意識障害 (Disturbance of consciousness)	36	281

Table 4: Examples of canonical terms with the largest number of variants in the hard- and soft-label dictionaries.

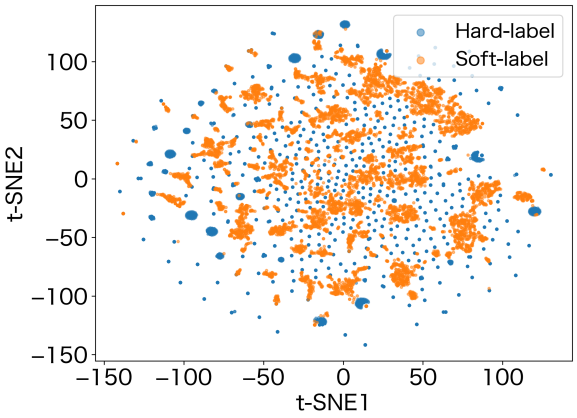

Figure 2: t-SNE visualization of vectorized label distributions for hard vs. soft dictionaries. Hard labels form distinct and isolated clusters due to their binary assignments, whereas soft labels yield broader and overlapping clusters that reflect probabilistic associations and linguistic ambiguity.

flects both the challenge of controlling fine-grained stylistic nuance in current LLMs and the naturally polite tone of patient-clinician communication.

4.4. Comparison with Existing Resources

We compared the JPPB with an existing Japanese resource, the patient expression dictionary (Nishidani et al., 2021). For coverage against ICD-10 (2013), we evaluated at the three-character level and considered a code covered if it was linked to at least one canonical term in the dictionary. Restricting to the R chapter (Symptoms and signs), the existing dictionary covers 146/276 codes (52.9%), whereas JPPB covers 172/276 codes (62.3%), an absolute gain of 9.4 percentage points. For reference, the existing dictionary's coverage across all ICD-10 chapters is 632/7,747 (8.2%); JPPB is designed for the R chapter and is not evaluated beyond it. Our resource covers a broader spectrum of symptom expressions and provides richer mappings, including multiple surface forms per canonical term and soft-label distributions. Unlike previous dictionaries that focused on one-to-one mappings at the word or short-phrase level, our phrase bank systematically captures the variability and ambiguity present in actual patient-generated language.

5. Evaluation

This section describes the experimental setup and the methods compared for evaluating the JPPB.

5.1. Experimental Setup

We evaluated the effectiveness of our resource using two Japanese medical NLP datasets: the KEEPHA Japanese Dataset (Raithel et al., 2024) and MedNLP-SC (Wakamiya et al., 2023). The KEEPHA dataset consists of tweets, whereas MedNLP-SC is a synthetic dataset. Both datasets contain texts describing symptoms. We prepro-

Dataset	Dictionary	Cosine Similarity	JS Distance	KL Divergence
KEEPHA	Hard label	0.085	0.928	20.968
	Soft label	**0.213**	**0.876**	**16.122**
MedNLP-SC	Hard label	0.111	0.908	20.502
	Soft label	**0.184**	**0.892**	**16.332**

Table 5: Comparison of label distribution similarity between hard- and soft-label dictionaries. Higher cosine similarity and lower divergence values indicate closer approximation to the gold label distribution. Best values for each dataset are shown in bold.

cessed each dataset by extracting symptom expressions corresponding to ICD-10 "R" codes, following the same filtering criteria used in our dictionary construction, and mapped them to canonical terms using our dictionaries.

For the KEEPHA dataset, which contains NER annotations, we extracted entities tagged as DISORDER and normalized them using JMED-DICT. Among these, we retained only the entities assigned ICD-10 codes beginning with "R." This procedure yielded a total of 40 instances.

For the MedNLP-SC corpus, in which 22 positive symptoms are annotated at the sentence level, we again restricted the set to ICD-10 "R"-category symptoms. Eleven such symptoms were included in this corpus: 悪心 (nausea), 倦怠感 (fatigue), 嘔気 (queasiness), 食欲不振 (loss of appetite), 頭痛 (headache), 発熱 (fever), めまい (dizziness), 疼痛 (pain), 過敏症 (hypersensitivity), 腹痛症 (abdominal pain), and 発疹 (rash). This filtering resulted in a dataset of 1,944 instances.

5.2. Methods and Baselines

We compared the performance of our hard-label and soft-label dictionaries with existing baselines, namely a large manually constructed patient expression dictionary, as well as with NER-based and LLM-based approaches.

For dictionary-based approaches, we report Top-k accuracy ($k = 1, 5, 10$) under an entry-based evaluation scheme. Specifically, the Top-k dictionary entries retrieved for a given patient expression were considered, and all canonical terms associated with those entries were treated as valid candidate labels. For hard labels, each entry contributed one canonical term, whereas for soft labels, each entry could contribute multiple canonical terms above the threshold. We adopted entry-based evaluation to ensure comparability with NER- and LLM-based methods, which typically return an unranked set of plausible canonical labels; thus, label-based Top-k metrics are not directly applicable to them. For NER- and LLM-based methods, we evaluated only their Top-1 prediction set.

5.2.1. Dictionary-based normalization

Patient-generated expressions were matched against entries in both the hard-label and soft-label versions of our phrase bank using embedding-based cosine similarity rather than exact string matching. Each input expression was encoded into a sentence embedding and compared with the embeddings of all patient-expression entries in the dictionary.

Hard-label: In the hard-label JPPB, each patient-expression entry was linked to a single canonical term. For each input expression, we retrieved the most similar dictionary entries based on cosine similarity and treated the canonical terms associated with those entries as candidate labels. Performance was evaluated using Top-1, Top-5, and Top-10 accuracy.

Soft-label: In the soft-label JPPB, each patient-expression entry was associated with multiple canonical terms, each assigned a score. For each input expression, we retrieved the most similar dictionary entries based on cosine similarity and expanded each entry into candidate canonical terms using the associated scores. Within each retrieved entry, only canonical terms with a score of at least 0.2 were retained and sorted by score. The final candidate list was constructed according to the rank order of the retrieved dictionary entries, and duplicate canonical terms were removed by keeping their first occurrence. Performance was evaluated using Top-1, Top-5, and Top-10 accuracy.

We additionally applied a **targeted setting**, in which the dictionaries were filtered to contain only canonical terms present in the evaluation dataset. This targeted setting should be interpreted as a constrained evaluation scenario, since the candidate space was restricted to canonical terms known to appear in the evaluation dataset. As a baseline, we evaluated an existing patient expression dictionary restricted to symptom entries (ICD-R codes).

5.2.2. NER-based normalization

We applied a BERT-based NER model (MedTXTNER[4]) to detect symptom spans in text. For normalization, each span s and each canonical term $c \in \mathcal{C}$ were encoded into sentence embeddings with the same encoder used in our resource construction. Cosine similarity $\cos(e_s, e_c)$ was computed, and the most similar canonical term was selected. JMED-DICT was used as the normalization dictionary.

5.2.3. LLM-based normalization

We prompted GPT-4.1-mini to directly output canonical symptom terms given a patient expression, without consulting any dictionary at inference time. The prompt instructed the model to choose the medically appropriate canonical terms. The generated terms were compared against the gold canonical terms to compute Top-1 accuracy. This setting isolates the LLM's ability to perform concept normalization without external lexical resources, serving as a strong reference point for dictionary-free deployment under our evaluation protocol.

Note that all methods compared in this study are unsupervised or zero-shot, as the goal is to evaluate resource-based normalization methods applicable in settings where task-specific labeled data are unavailable. Supervised classification models, such as fine-tuned BERT variants, were therefore not included.

6. Results

Table 6 summarizes performance across datasets and methods. Three main observations can be made.

Soft vs. Hard Labels: On the KEEPHA dataset, the soft-label dictionary achieved higher Top-1 accuracy than the hard-label dictionary (0.300 vs. 0.100), and also higher Top-5 accuracy (0.475 vs. 0.350). For Top-10, the difference between soft and hard labels was smaller (0.500 vs. 0.425). A similar trend was observed for the MedNLP-SC dataset, where soft labels reached 0.257 Top-1 accuracy compared to 0.134 for hard labels, and 0.568 Top-10 accuracy compared to 0.421.

Baseline vs. Soft Labels: Compared with the baseline, soft labels did not consistently outperform across all metrics. On KEEPHA, the baseline dictionary achieved the highest Top-10 accuracy (0.525), while soft labels reached higher Top-1 accuracy (0.300 vs. 0.225). On MedNLP-SC, soft labels slightly outperformed the baseline in both Top-1 accuracy (0.257 vs. 0.256) and Top-10 accuracy (0.568 vs. 0.556).

Comparison with NER- and LLM-based Methods: NER- and LLM-based methods were included for reference. On KEEPHA, LLM-based normalization attained the highest Top-1 accuracy (0.575) and targeted LLM further improved to 0.953. On MedNLP-SC, targeted NER achieved 0.760 Top-1 accuracy and 0.530 F1, both higher than dictionary-based methods. These methods generally outperformed dictionaries in accuracy and F1 under Top-1 evaluation.

6.1. Error Analysis

We conducted a detailed error analysis on the KEEPHA dataset, comparing the Top-20 entry-based normalization results of the soft-label dictionary and the existing baseline for each expression.

Soft vs. Baseline: To further examine the differences between the soft-label dictionary and the baseline, we conducted a paired error analysis. The majority of cases (90%) fell into the *both-hit* category, indicating that the two methods are simultaneously correct for most patient expressions. Only a small fraction of cases were classified as *soft-only hit* (5%), *existing-only hit* (2.5%), or *both miss* (2.5%).

Some soft-only cases appear to result from the annotation process. For example, the expression お腹痛い ("My stomach hurts") was annotated only as 疼痛 (pain) despite its clear connection to 腹痛症 (abdominal pain) or 胃痛 (stomachache). Other examples such as めまい (dizziness) and 吐き気 (nausea) show that soft labels partially captured one gold term, 嘔気 (nausea), whereas the existing dictionary failed.

In existing-only cases, the errors of the soft-label dictionary were clear mismatches. For example, in the sentence "シクロフォスファミドの副作用…浮腫（血管内に水が残っている）…" ("Side effects of cyclophosphamide…edema (water remaining in blood vessels)…"), the gold labels included 浮腫 (edema), which the existing dictionary correctly retrieved. In contrast, the soft-label dictionary produced unrelated terms such as 腹痛症 (abdominal pain), 胃痛 (stomachache), and 疼痛 (pain), reflecting a failure in precision rather than broad coverage. This highlights that while soft labels improve recall for colloquial or broad patient expressions, they can also introduce incorrect mappings for precise medical terms.

Soft vs. Hard Labels: We also compared the proposed soft-label dictionary with its hard-label counterpart. The two methods overlapped on most cases (*both-hit*: 85%), indicating similar performance for the majority of expressions. Approximately 10% of the instances were *soft-only hit*, while no cases were observed where hard labels

[4] https://huggingface.co/sociocom/MedTXTNER

Dataset	Setting	Approach	Method	Accuracy			F1		
				Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
KEEPHA	Untargeted	Dictionary	Baseline	0.225	0.400	**0.525**	**0.200**	**0.244**	**0.202**
			Hard Label	0.100	0.350	0.425	0.083	0.166	0.136
			Soft Label	**0.300**	**0.475**	0.500	0.195	0.156	0.138
		Model	NER	0.186	—	—	0.271	—	—
			LLM	0.575	—	—	0.414	—	—
	Targeted	Dictionary	Baseline	**0.500**	**0.800**	**0.850**	**0.441**	**0.516**	**0.437**
			Hard Label	0.325	0.625	0.825	0.283	0.383	0.430
			Soft Label	0.450	0.675	0.825	0.378	0.388	0.396
		Model	NER	0.535	—	—	0.386	—	—
			LLM	0.953	—	—	0.818	—	—
MedNLP-SC	Untargeted	Dictionary	Baseline	0.256	**0.458**	0.556	**0.201**	**0.223**	**0.192**
			Hard Label	0.134	0.325	0.421	0.107	0.145	0.128
			Soft Label	**0.257**	0.449	**0.568**	0.160	0.149	0.135
		Model	NER	0.300	—	—	0.302	—	—
			LLM	0.565	—	—	0.373	—	—
	Targeted	Dictionary	Baseline	**0.532**	0.690	0.764	**0.431**	**0.448**	**0.431**
			Hard Label	0.443	**0.710**	**0.802**	0.364	0.418	0.413
			Soft Label	0.490	0.707	0.791	0.388	0.427	0.424
		Model	NER	0.760	—	—	0.530	—	—
			LLM	0.689	—	—	0.551	—	—

Table 6: Normalization performance on the KEEPHA and MedNLP-SC datasets. We report Top-k accuracy (Top-1, Top-5, and Top-10) and F1 scores for the dictionary-based methods (baseline, hard-label, and soft-label) under both untargeted and targeted settings. NER- and LLM-based normalization are shown as references, with Top-1 accuracy and F1 only. Soft labels generally outperform hard labels in Top-1 accuracy, although hard labels perform slightly better in some settings. The baseline remains strong, and targeted settings improve all methods. Bold values indicate the best scores for each dataset and metric.

succeeded and soft labels failed. This suggests that soft labels can improve recall and coverage in ambiguous cases, although the gains were not uniform across all evaluation metrics.

Qualitatively, soft labels better captured broad or colloquial expressions such as 倦怠感 (fatigue), 発熱 (fever), and 腹痛症 (abdominal pain), whereas hard labels were more prone to mismatches in ambiguous contexts. For example, in the sentence “シスプラチンの副作用は倦怠感、吐き気、高熱…” (“Side effects of cisplatin include fatigue, nausea, and high fever…”), the gold labels included 倦怠感 (fatigue) and 嘔気 (nausea), which were successfully retrieved by the soft dictionary but missed in the hard setting. Conversely, both methods failed on highly specific expressions such as 不整脈 (arrhythmia) and 浮腫 (edema), suggesting that dictionary coverage still limits precision for narrowly defined medical terms.

7. Discussion

The comparison between hard-label and soft-label dictionaries indicates that soft labels provide better coverage in several settings, particularly for Top-1 accuracy. This suggests that allowing multiple canonical candidates can help capture ambiguity in patient expressions, although the advantage was not uniform across all Top-k and F1 metrics.

When comparing the soft-label dictionary with the baseline, results show that the baseline tends to achieve higher accuracy, especially for Top-10. However, soft labels often yield higher recall, highlighting their potential advantage in capturing broader sets of plausible medical terms. This trade-off suggests that soft labels suit recall-oriented tasks, whereas the baseline favors precision.

Our results show that the automatically constructed soft-label dictionary achieves performance levels comparable to manually curated patient expression resources. While the handcrafted baseline remains slightly stronger overall, the gap is narrow, especially in recall-oriented settings. This suggests that large-scale, labor-intensive manual curation may not always be necessary: with our approach, dictionaries of practical utility can be built automatically, offering a scalable alternative for languages and domains where expert annotation is costly or unavailable.

In contrast, LLM-based methods outperformed

dictionary-based approaches in Top-1 accuracy and F1, especially under targeted settings. These models can flexibly generate candidate terms without requiring task-specific annotation, but they demand substantial computational resources and are often deployed through commercial APIs. Such API-based usage raises practical challenges in healthcare environments, where privacy regulations, data governance policies, and internet connectivity constraints restrict the transfer of sensitive text data. This highlights the practicality of dictionary-based resources like JPPB, which are lightweight, transparent, and locally deployable.

We therefore regard LLM-based results as an upper-bound reference rather than a directly applicable solution. Importantly, the two approaches are not mutually exclusive: under conditions where privacy and deployment constraints can be appropriately managed, hybrid strategies may enable dictionaries to constrain or post-process LLM outputs, thereby combining the performance benefits of LLMs with the stability and practicality of dictionary-based methods.

8. Conclusion and Future Work

In this study, we presented the construction and comprehensive evaluation of the JPPB, a large-scale resource mapping diverse patient-generated expressions to standardized symptom terminology. Unlike conventional manually curated lexicons, JPPB was built through a semi-automatic pipeline combining persona-based expression generation and soft label assignment, enabling scalable construction and continuous extension. While its performance did not consistently exceed that of existing dictionaries, the automated design allows JPPB to capture linguistic variability and uncertainty more flexibly, offering a practical foundation for future normalization and retrieval tasks.

Looking ahead, future work includes the continuous expansion and refinement of the phrase bank, particularly by incorporating new patient expressions from real-world clinical narratives and emerging health topics. Further improvements can be made by integrating contextual information, such as patient demographics or clinical histories, to resolve remaining ambiguities. We also plan to explore its applications as a benchmark for the development and evaluation of advanced machine learning models in Japanese medical NLP.

9. Limitations

A limitation of our current design is that soft labels are derived by propagating from hard-label assignments via embedding similarity. This procedure risks inheriting and amplifying the biases of the hard-label construction, and thus the resulting distributions should be regarded as an approximation for linguistic ambiguity rather than ground-truth labels. Nevertheless, because each hard label ultimately derives from dictionary-based canonical terms, which are already paired with diverse surface forms, the propagated soft labels are not arbitrary. Instead, they provide a heuristic mechanism to extend one-to-one mappings into broader sets of plausible associations. In this way, our approach leverages the observed diversity of patient expressions while remaining grounded in medically valid terminology. Future work could mitigate remaining biases by incorporating human annotation of multiple labels or by directly leveraging LLMs to generate diverse candidate mappings.

10. Acknowledgments

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP), Project JPJ012425, and the Japan Society for the Promotion of Science (JSPS) Research Start-up Grant JP25K24412.

11. Bibliographical References

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Gen Gu, Xingting Zhang, Xingeng Zhu, Zhe Jian, Ken Chen, Dong Wen, Li Gao, Shaodian Zhang, Fei Wang, Handong Ma, and Jianbo Lei. 2019. Development of a Consumer Health Vocabulary by Mining Health Forum Texts Based on Word Embedding: Semiautomatic Approach. *JMIR Medical Informatics*, 7(2):e12704.

E. Hayes, R. Dua, E. Yeung, and K. Fan. 2017. Patient understanding of commonly used oral medicine terminology. *British Dental Journal*, 223(11):842–845.

Mohammed Ibrahim, Susan Gauch, Omar Salman, and Mohammed Alqahtani. 2021. An

automated method to enrich consumer health vocabularies using GloVe word embeddings and an auxiliary lexical resource. *PeerJ. Computer Science*, 7:e668.

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao, and Eiji Aramaki. 2018. J-MeDic: A Japanese disease name dictionary based on real clinical usage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Alla Keselman, Catherine Arnott Smith, Guy Divita, Hyeoneui Kim, Allen C. Browne, Gondy Leroy, and Qing Zeng-Treitler. 2008. Consumer health concepts that do not map to the UMLS: where do they fit? *Journal of the American Medical Informatics Association: JAMIA*, 15(4):496–505.

Mihiro Nishidani, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. 生成アプローチによる患者表現の標準化 [standardization of patient expressions via a generative approach] (in japanese). In *Proceedings of the JSAI Special Interest Group on Artificial Intelligence in Medicine (SIG-AIMED)*. The Japanese Society for Artificial Intelligence.

Tomohiro Nishiyama, Ayane Yamaguchi, Peitao Han, Lis Weiji Kanashiro Pereira, Yuka Otsuki, Gabriel Herman Bernardim Andrade, Noriko Kudo, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki, Masahiro Takada, and Masakazu Toi. 2024. Automated System to Capture Patient Symptoms From Multitype Japanese Clinical Texts: Retrospective Study. *JMIR Medical Informatics*, 12:e58977.

Yukiko Ohno, Riri Kato, Haruki Ishikawa, Tomohiro Nishiyama, Minae Isawa, Mayumi Mochizuki, Eiji Aramaki, and Tohru Aomori. 2024. Using the Natural Language Processing System Medical Named Entity Recognition-Japanese to Analyze Pharmaceutical Care Records: Natural Language Processing Analysis. *JMIR Formative Research*, 8:e55798.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Abeed Sarker, Rui Zhang, Yanshan Wang, Yunyu Xiao, Sudeshna Das, Dalton Schutte, David Oniani, Qianqian Xie, and Hua Xu. 2024. Natural Language Processing for Digital Health in the Era of Large Language Models. *Yearbook of Medical Informatics*, 33(01):229–240.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629, Miami, Florida, USA. Association for Computational Linguistics.

Q. Zeng, S. Kogan, N. Ash, R. A. Greenes, and A. A. Boxwala. 2002. Characteristics of Consumer Terminology for Health Information Retrieval. *Methods of Information in Medicine*, 41(04):289–298.

Qing T. Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association: JAMIA*, 13(1):24–29.

Zeng Qing, Kogan Sandra, Ash Nachman, and Greenes Robert A. 2001. Patient and Clinician Vocabulary: How Different Are They? In *Studies in Health Technology and Informatics*. IOS Press.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When ”A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Maryam Zolnoori, Kin Wah Fung, Timothy B. Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D. Shah, Yi Shuan Shirley Wu, Christina E. Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in Brief*, 24:103838.

12. Language Resource References

Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol,

Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. A dataset for pharmacovigilance in German, French, and Japanese: Annotating adverse drug reactions across languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 395–414, Torino, Italia. ELRA and ICCL.

Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Katherine Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview.

13. Appendix

This appendix provides additional experimental results, prompt design details, and error analysis to complement the main findings of this study.

13.1. Prompt Design for Expression Generation

The original Japanese prompt and its English translation are shown in Figures 3 and 4, respectively. These figures illustrate the prompt template used to generate patient expressions from canonical medical terms. The prompt is designed to encourage natural, non-medical language while maintaining semantic relevance to the given symptom. Specifically, it instructs the model to describe symptoms from a first-person perspective, avoid medical terminology, and provide multiple variations. The prompt was implemented in a role-based format, where a system message was used to specify the patient persona (e.g., elderly), and the instruction prompt was provided as the user message.

13.1.1. System Prompts

We used the following system prompts to specify patient personas:

- あなたは日本の一般人です。(*You are a general adult in Japan.*)
- あなたは日本人の子どもです。(*You are a child in Japan.*)
- あなたは日本の高齢者です。(*You are an elderly individual in Japan.*)

 # Instructions
あなたは病院の外来を受診した患者です。以下の「出現形」と「正規形」（医療用語）を参考に、医師に自分の症状を自然な言葉で伝えてください。

- 「どこが」「どんなとき」「どんなふうに」「どんなきっかけで」など、症状の場所、状況、感じ方も自由に加えてください。
- 誰かの話ではなく、自分の症状として話してください。
- 医療用語は使わず、家族や友人に話すような言い回しにしてください。
- ３パターン作成してください。
- 出現形・正規形に関連する症状のみに言及し、それ以外の症状には触れないでください。

 # Example
Input: 出現形: できものができている / 正規形: 皮下腫瘤

Output:

- 腕に小さいできものができて、押すとちょっと痛いです
- 首の後ろにしこりみたいなものができて、不安です

Figure 3: Japanese Prompt template used for generating patient expressions from canonical medical terms.

13.2. Additional Analysis of Persona Differences

Table 7 presents descriptive statistics of sentence length across different personas. The results show that sentence lengths are broadly comparable across personas, with only minor differences in mean and variance.

Persona	Mean	Std	Median	25%	75%
Child	41.10	9.56	40	34	47
Elderly	43.38	10.91	42	35	50
General	42.20	9.97	41	35	48

Table 7: Descriptive statistics of sentence length across different personas.

13.3. Precision and Recall Analysis

Table 8 presents a detailed comparison of precision and recall across datasets and methods. Overall, the baseline dictionary achieves the highest precision in most settings, particularly for Top-k

 # Instructions
You are a patient visiting an outpatient clinic. Based on the given “surface form” and “canonical form” (medical term), describe your symptoms to a doctor in natural language.

- You may include details such as where the symptom occurs, when it happens, how it feels, and what triggered it.
- Describe the symptoms as your own experience.
- Do not use medical terminology; use expressions as if speaking to family or friends.
- Generate three variations.
- **Only mention symptoms related to the given forms. Do not include unrelated symptoms.**

 # Example
Input: Surface form: lump / Canonical form: subcutaneous tumor

Output:

- I have a small lump on my arm, and it hurts when I press it.
- There is something like a lump on the back of my neck, and I am a bit worried.

Figure 4: English translation of the prompt template shown in Figure 3.

evaluation, indicating strong performance in exact matching scenarios.

In contrast, the proposed soft-label dictionary consistently improves recall compared to the hard-label setting, especially in Top-1 and Top-5 metrics. This suggests that allowing multiple candidate labels per expression enhances coverage and better captures the variability of patient language.

In targeted settings, all methods show substantial improvements, as the candidate space is restricted to relevant concepts. Notably, the soft-label approach achieves the highest recall in several conditions, demonstrating its effectiveness in capturing ambiguous or broad patient expressions.

Dataset	Method	Precision			Recall		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
KEEPHA	Baseline	**0.225**	**0.213**	**0.141**	0.188	0.338	**0.421**
	Hard Label	0.100	0.136	0.089	0.075	0.275	0.350
	Soft Label	0.175	0.108	0.093	**0.250**	**0.368**	0.405
	Targeted						
	Baseline	**0.500**	**0.473**	**0.355**	0.413	0.634	0.709
	Hard Label	0.325	0.344	**0.359**	0.263	0.501	0.697
	Soft Label	0.392	0.334	0.318	**0.388**	**0.555**	**0.718**
MedNLP-SC	Baseline	**0.256**	**0.203**	**0.140**	0.176	0.329	0.404
	Hard Label	0.134	0.120	0.089	0.095	0.245	0.317
	Soft Label	0.169	0.112	0.092	**0.177**	**0.331**	**0.435**
	Targeted						
	Baseline	**0.533**	**0.465**	**0.397**	0.385	0.515	0.600
	Hard Label	0.443	0.394	0.359	0.330	**0.554**	0.651
	Soft Label	0.457	0.403	0.369	**0.365**	0.556	**0.640**

Table 8: Precision and Recall comparison across datasets and methods. Best values for each dataset and metric are shown in bold.