Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# StrokeID-NER : A Stroke Named Entity Recognition Dataset for Indonesian Patient- Generated Health Queries

**Miseri Cordiaz S**
**Yaseen Muhammad**
**Md Ariful Islam Mozumder**
**Hee Cheol Kim**
heeki@inje.ac.kr

**Article**

**Keywords:**

**Posted Date:** July 27th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10163348/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# StrokeID-NER : A Stroke Named Entity Recognition Dataset for Indonesian Patient-Generated Health Queries

**Miseri Cordiaz S[1], Muhammad Yaseen[1], Md Ariful Islam Mozumder[1], and Hee-Cheol Kim[1,*]**

[1]Department of Digital Anti-Aging Healthcare, Inje University, Gimhae-si, 50834, Republic of Korea
[*]heeki@inje.ac.kr

## ABSTRACT

Stroke is the leading cause of death and disability in Indonesia, yet clinical NLP tools for stroke triage support remain limited by the absence of domain-specific annotated resources for patient-generated health text. We introduce StrokeID-NER, a stroke-specific named entity recognition (NER) dataset for informal Indonesian patient queries, comprising 1,742 queries annotated with 6,765 entity spans across four clinically grounded entity types — Symptom, Diagnosis, Temporal, and Risk Factor — with inter-annotator agreement ($\kappa$ = 0.857). We fine-tuned and evaluated five NER systems on our corpus, spanning monolingual and multilingual transformer models at two scales and a GPT-5.4 zero-shot baseline. Fine-tuned XLM-RoBERTa-large achieved the best performance (macro F1 = 0.7823), outperforming the zero-shot baseline by 11.4 points, with the largest advantage concentrated in Symptom and Risk Factor entities requiring recognition of informal Indonesian, Javanese regional terminology, and clinical context. Hapax rates of 72.0–86.9% across entity types and 68.3% cross-model false negative overlap indicate that performance gains are constrained by lexical coverage rather than model architecture, suggesting vocabulary expansion as the most effective path to improvement. StrokeID-NER, the annotation guidelines, and all trained models are released publicly at https://doi.org/10.5281/zenodo.20305007.

## Introduction

Stroke is a leading cause of death and disability in Indonesia[1]. Indonesia has the highest age and sex standardized mortality rate in Southeast Asia. It also shows the largest loss of disability-adjusted life years[2]. In the region, Indonesia ranks second in stroke incidence and prevalence. It reports 642,943 new cases and 4,918,487 people living with stroke[3]. Telemedicine for stroke patients offers a promising approach for effective acute stroke management[1].

Digital health consultation platforms are now a keyway for Indonesian patients to get medical advice. Alodokter.com has built a large archive of health questions from patients in casual Indonesian. These questions are a valuable but unstructured resource. Patients and caregivers describe symptoms, risk factors, and medical histories in simple terms, without formal medical language. Extracting structured clinical information from this patient-generated data can help support clinical decision systems, triage tools, and health surveillance at the population level[4]. Named entity recognition (NER), the task of identifying and classifying clinically relevant spans of text, is a foundational step enabling these downstream applications[5].

Named entity recognition (NER) is an essential subtask in natural language processing (NLP), and it still suffers from data sparseness for the majority of languages, including Indonesian[6]. Despite being spoken by over 200 million people and ranking among the world's ten most widely spoken languages, Indonesian remains significantly underrepresented in NLP research[7–9]. A unique challenge in developing Indonesian NER compared to other languages is that Indonesian has many words that can have multiple meanings or ambiguous contexts and also more flexible sentence structure (with more use of phrases or conjunctions) makes the NER task more challenging as entities can be in different positions in the sentence[10].

Prior work on Indonesian medical NER has made initial progress but remains limited in scope. Kusumawardani and Kusumawati[11] developed a health consultation NER system using a BiLSTM-CRF architecture, combining patient questions and physician answers from three medical consultation platforms - Alodokter, Klikdokter, and SehatQ. Ignasius et al.[12] constructed the DetikHealth corpus from Indonesian health-news articles, benchmarking Word2Vec and IndoBERT embeddings with BiLSTM-CRF models across three general medical entity types (disease, symptom, drug). Both studies confirm the value of domain-specific annotation for Indonesian medical NLP, yet neither addresses stroke as a clinical domain, and neither focuses exclusively on patient-generated informal queries. Stroke queries represent a distinct NLP challenge: the combination of time-critical clinical content, extreme orthographic variation in informal Indonesian, Javanese and regional lay terminology, and Islamic cultural temporal references cannot be adequately addressed by existing Indonesian NLP resources or by NLP systems developed for formal clinical text in other languages.

This work addresses this gap by introducing StrokeID-NER, a stroke-specific NER dataset for Indonesian patient-generated health queries. The dataset comprises 1,742 patient queries from alodokter.com, annotated with 6,765 entity spans across four clinically grounded entity types : Symptom (SYM), Diagnosis (DGN), Temporal (TMP), and Risk Factor (RF). Entity definitions are grounded in established clinical frameworks including the AHA/ASA stroke management guidelines[13], Indonesian National Clinical Practice Guidelines for Stroke[14], ICD-10 cerebrovascular disease classifications[15], and Riskesdas 2018 national health data[16]. We benchmark five NER systems on StrokeID-NER: four fine-tuned transformer models spanning monolingual Indonesian and multilingual pretraining strategies at two model scales, and a GPT-5.4 zero-shot baseline. This work makes four contributions :

- It releases the first publicly available, clinically grounded corpus for stroke named entity recognition in Indonesian
- It benchmarks how pretraining language, model scale, and decoding strategy affect performance on informal Indonesian clinical NER
- It provides a detailed linguistic analysis of patient-generated stroke text, documenting orthographic variation, dialectal terminology, code-mixing, and culturally specific temporal expressions
- It demonstrates empirically that fine-tuned multilingual models consistently and substantially outperform zero-shot large language model prompting for this task

## Methodology

StrokeID-NER construction and the benchmarking of five NER systems follow four phases: data collection and filtering, corpus annotation and splitting, model training, and evaluation. Figure 1 summarizes this sequence. Each phase receives detailed treatment in subsequent subsections, establishing the experimental framework for testing all systems on the final corpus.

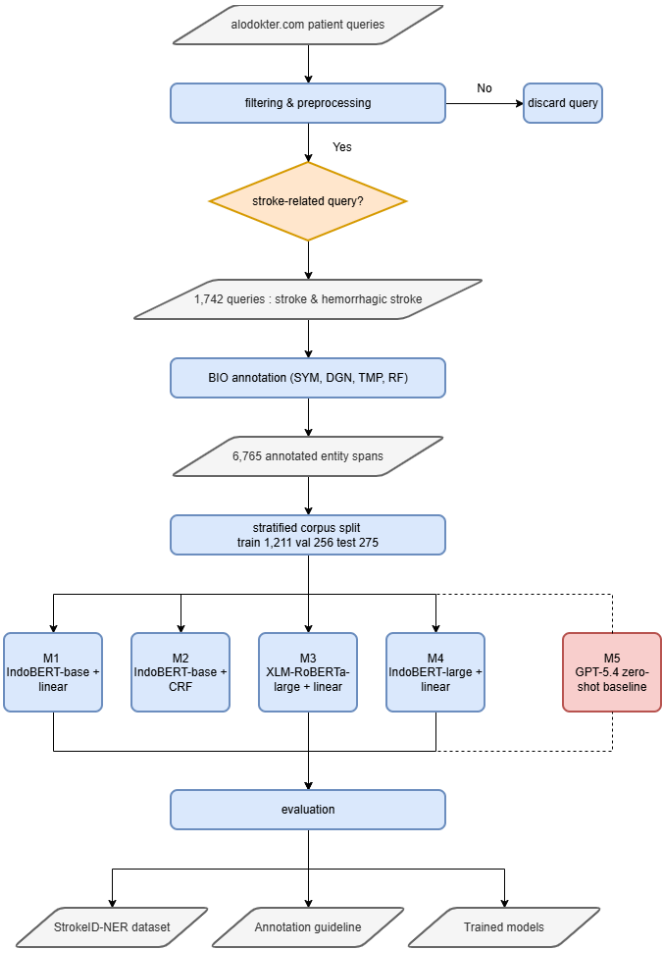



**Figure 1.** Research workflow for StrokeID-NER.

## Data Source

The Indonesia-medical-qna dataset[17] on HuggingFace, which was gathered from alodokter.com, the top digital health consultation platform in Indonesia, was used to build StrokeID-NER. By connecting over 30 million monthly active users with a network of over 80,000 licensed physicians, Alodokter creates a vast archive of naturally occurring patient-generated health queries in Indonesian and enables patients and caregivers to ask doctors questions. After preprocessing to eliminate duplicates and non-contextual content, we filtered queries tagged under stroke and hemorrhagic stroke topics, resulting in 1,742 patient queries.

StrokeID-NER contains exclusively patient and caregiver submitted queries, deliberately excluding physician responses. A clinical triage support system functions on patient input alone. Physician responses would introduce formal, structured, clinically consistent language that fails to reflect deployment conditions. This approach differs from prior Indonesian medical NER work[11], which merged patient questions and physician answers into one corpus. Beyond age references retained as clinically relevant Risk Factor (RF) annotations the dataset contains no personally identifiable information. All data remains publicly accessible through the original HuggingFace repository.

## Annotation Scheme

We define four entity types grounded in established clinical frameworks for stroke assessment and triage. All entities are annotated in BIO (Begin-Inside-Outside) format, yielding a label set of nine tags: {O, B-SYM, I-SYM, B-DGN, I-DGN, B-TMP, I-TMP, B-RF, I-RF}. Table 1 summarizes the entity definitions, clinical grounding, and representative span examples. Entity boundaries were validated by a faculty advisor with expertise in digital healthcare and evaluated by doctoral students for methodological consistency. The complete annotation guideline is released alongside the dataset.

| Entity | Definition | Clinical Grounding | Example Spans |
|---|---|---|---|
| SYM (Symptom) | Physical or neurological symptoms reported as a consequence of stroke, including formal clinical terms and informal patient expressions | AHA/ASA FAST criteria[13,18]; PERDOSSI 2019 | *lemas* (weakness), *drodog* (tremor), *terasa seperti kesetrum* (feels like electrocuted) |
| DGN (Diagnosis) | Stroke diagnoses, subtypes, and related cerebrovascular conditions | ICD-10 codes I60-I69[15]; PERDOSSI 2019 | *stroke ringan* (mild stroke/TIA), *pendarahan otak* (brain hemorrhage) |
| TMP (Temporal) | Temporal expressions indicating onset, duration, or timing of symptoms or diagnoses | Original contribution; not formalized in existing stroke guidelines | *tiba-tiba* (suddenly), *sudah 3 hari* (for 3 days), *saat hari ke-20 puasa* (20th day of Ramadhan) |
| RF (Risk Factor) | Pre-existing conditions, demographics, and lifestyle factors that increase stroke risk | PERDOSSI 2019; RISKESDAS 2018 | *hipertensi*, *diabetes*, *usia 65 tahun* (age 65 years) |

**Table 1.** StrokeID-NER entity types definitions

## Annotation Process

Annotation was conducted using Label Studio[19], an open-source data labeling platform, by a team of four members: a primary annotator who labeled all 1,742 queries; a second annotator (public health student) who independently labeled a stratified random sample of 100 queries for inter-annotator agreement assessment; a faculty advisor who validated entity definitions for clinical appropriateness; and a doctoral student who evaluated methodological consistency and linguistic coverage.

The annotation guideline was developed iteratively. An initial set four core rules were established prior to annotation, grounded in the clinical references described above. Annotation guidelines expanded from four to nineteen rules as edge cases surfaced during progressive review. Stroke-related diagnoses were separated from other neurological conditions. Negated entities like *tidak ada hipertensi* received O labels, effectively becoming non-entities. Javanese colloquialisms (*drodog*, *cekot-cekot*, *kolep*) qualified as valid SYM spans when their clinical meanings were clear. Medical procedures such as CT scans and MRI fell outside annotation scope, as did recovery narratives. Comparative queries lacking reported symptoms were classified as zero-entity tasks.

## Inter-Annotator Agreement

A stratified random sample of 100 queries, balanced across entity types to guarantee representation of all four entity categories, was independently annotated by the second annotator to evaluate annotation quality. In order to ensure independent annotation, no reference to the labels of the primary annotator was made during this process. As is customary in NER annotation studies, Cohen's Kappa[20] was used to measure inter-annotator agreement at the token level. The overall token-level Cohen's Kappa was $\kappa = 0.857$, exceeding the widely recognized threshold of $\kappa \geq 0.80$ for high-quality NLP annotation and indicating nearly perfect agreement[21]. The primary annotator adjudicated all disputes in accordance with the established guidelines, and the adjudicated annotation was kept as the primary benchmark.

## Dataset Statistics

The full dataset statistics are shown in Table 2 with SYM is the most common entity type (39.1%), which reflects the symptom-centric nature of patient health consultation queries. The corpus comprises 6,765 annotated entity spans across 1,742 queries. To ensure repeatable results and balanced entity distributions across splits, the dataset was divided into training (1,211 queries, 69.5%), validation (256 queries, 14.7%), and test (275 queries, 15.8%) splits using stratified sampling with a fixed random seed (42).

| Entity | Train | Val | Test | Total | Unique | Hapax | Hapax% |
|---|---|---|---|---|---|---|---|
| SYM | 1,816 | 391 | 436 | 2,643 | 1,326 | 1,061 | 80.0% |
| DGN | 1,527 | 308 | 354 | 2,189 | 282 | 203 | 72.0% |
| TMP | 756 | 167 | 171 | 1,094 | 693 | 571 | 82.4% |
| RF | 583 | 122 | 134 | 839 | 496 | 431 | 86.9% |
| **Total** | **4,682** | **988** | **1,095** | **6,765** | **2,788** | — | — |

**Table 2.** StrokeID-NER dataset statistics

The high hapax rate, the percentage of unique entity expressions that appear only once in the corpus, ranging from 72.0% for DGN to 86.9% for RF is one of the dataset's distinguishing features. These rates are significantly higher than those found in clinical NER datasets derived from official medical records, where lexical diversity is reduced due to standardized clinical terminology[22]. As mentioned in Section 3.7, the high hapax rates pose a serious problem for NER models and reflect the informal, dialectal, and highly variable nature of patient-generated text. SYM has the highest absolute hapax count (1,061) as well as the largest unique expression count (1,326), which reflects the striking diversity in how patients describe stroke symptoms, ranging from inventive lay metaphors and regional expressions recorded in Section 2.6 to standard clinical terms (*lemas*, weakness; *pusing*, dizziness). Despite having the second-largest span count, DGN has the lowest unique count (282).

## Linguistic Characteristics

Five unique linguistic phenomena found in patient-generated stroke queries in Indonesian fundamentally set this corpus apart from formal clinical text and pose special difficulties for NER systems.

*Orthographic variation*

There are more than 20 phonetic variations of the word stroke alone (*struk*, *strok*, *setruk*, *sroke*). Similar differences can be observed in multi-word entities: hypertension is expressed as *hipertensi*, *darah tinggi*, and *tekanan darah tinggi* (153 occurrences across three surface forms), whereas sudden onset is expressed as *tiba-tiba*, *tiba2*, and *mendadak* (182 occurrences across four forms).

*Metaphor mapping from lay to clinical*

Patients describe clinical phenomena using everyday metaphors that have no lexical similarity to their clinical equivalents: gait imbalance is described as *terasa seperti menggunakan* high heels *sebelah* (feels like wearing a high heel on one side); paresthesia is described as *terasa seperti kesetrum* (feels like being electrocuted) or *serasa darahnya ngak ngaliiiir* (feels like the blood is not flowing); and facial numbness is described as *pipi seperti ngemut permen* (cheek feels like holding candy). The structural performance ceiling represented by these single-occurrence hapax expressions cannot be addressed by architectural advancements alone.

*Javanese and local colloquialisms*

In Javanese and other regional languages, symptoms include *drodog* (tremor), *kolep* (collapse), *cekot-cekot* (throbbing pain), and *klenger* (extreme weakness). These terms are absent from standard Indonesian NLP resources, and models trained on formal text are unable to recognize them.

*Linguistic hybridization and code-mixing*

The corpus includes multilingual code-mixing in Indonesian, Javanese, English, and French, as well as one French term (*AVC cérébral*) and English medical abbreviations (TIA, ICH, and CVA). The use of XLM-RoBERTa, whose multilingual pretraining offers richer subword representations for code-mixed text compared to monolingual Indonesian models, is directly motivated by this.

*Culturally specific temporal expressions*

Temporal references include Islamic calendar expressions absent from English clinical datasets: *saat hari ke-20 puasa* (20th day of Ramadhan fasting), *habis sahur* (after pre-dawn Ramadhan meal), and *menjelang maghrib* (approaching evening prayer time).

All of these traits support the idea that patient-generated stroke text in Indonesian is a unique area of natural language processing. NLP systems designed for formal text cannot effectively handle the high hapax rates (72.0–86.9%), orthographic variation, lay metaphors, regional terminology, and cultural temporal references. This highlights the need for a specialized annotated corpus like StrokeID-NER.

## Task Formulation

Under the BIO tagging scheme, we formulate stroke NER as a sequence labeling task, giving each token a label from Y = {O, B-SYM, I-SYM, B-DGN, I-DGN, B-TMP, I-TMP, B-RF, I-RF}. Figure 2 uses an example patient query to show the entire annotation and inference pipeline.

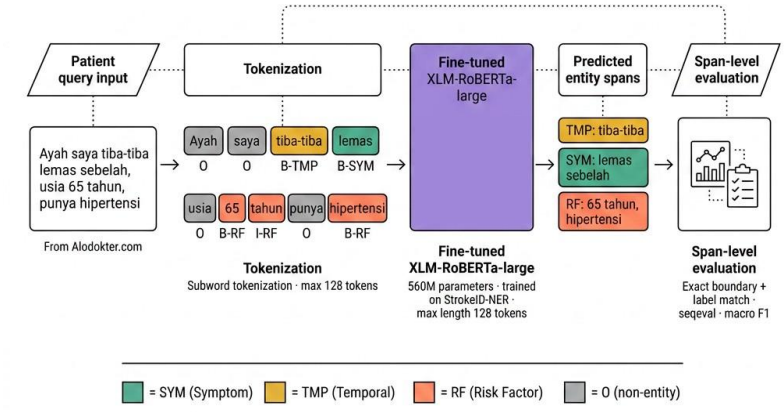



**Figure 2.** BIO annotation and inference pipeline for StrokeID-NER.

Only when a predicted span's token boundaries and entity label precisely match the gold annotation is it considered correct; partial boundary matches are not given credit. In accordance with the standard NER evaluation procedure established by CoNLL-2003[23], evaluation is carried out using seqeval[24].

## Models

In order to enable controlled comparisons across these dimensions, we assess five systems covering a variety of pretraining strategies, model scales, decoding mechanisms, and prompting approaches.

*IndoBERT-base + Linear (M1)*

We fine-tune IndoBERT-base-p2[25], a BERT-base architecture pretrained on 23.43 GB of Indonesian text. A linear classification head projects from 768-dimensional token representations to scores over 9 BIO labels. This model establishes the performance baseline for monolingual Indonesian fine-tuning.

*IndoBERT-base + CRF (M2)*

M1 augmented with a Conditional Random Field (CRF) layer[26] placed above the linear projection, enforcing globally consistent BIO label sequences by preventing invalid transitions such as I-SYM following B-DGN. Comparing M1 and M2 isolates the contribution of structured sequential decoding while holding all other components constant.

*XLM-RoBERTa-large + Linear (M3)*

We fine-tune the cahya/xlm-roberta-large-indonesian-NER checkpoint[27], an XLM-RoBERTa-large model[28] additionally pretrained on Indonesian NER data. A linear head is applied to 1,024-dimensional token representations. This model represents our primary multilingual hypothesis: that cross-lingual pretraining on 100 languages produces richer subword representations for informal, code-mixed, and regionally varied Indonesian text than monolingual pretraining alone.

*IndoBERT-large + Linear (M4)*

The large variant of IndoBERT[25] with 334M parameters and 1,024-dimensional hidden states with a linear head. Comparing M1 and M4 isolates model scale within the monolingual family; comparing M4 and M3 isolates multilingual pretraining while controlling for scale, as both use 1,024-dimensional architectures.

*GPT-5.4 — Zero-Shot Baseline*

GPT-5.4[29] is evaluated in a zero-shot setting with a structured system prompt specifying entity definitions, annotation rules, and BIO labeling conventions. The model returns a JSON array of token-label pairs. No fine-tuning, few-shot examples, or chain-of-thought reasoning are applied. This baseline quantifies the value of domain-specific annotation against a state-of-the-art general-purpose LLM. GPT-5.4 was used solely as an evaluated baseline system in this study and was not used in the preparation or writing of this manuscript.

### Training Configuration

All fine-tuned models are trained on the training split (n = 1,211) with the best checkpoint selected on macro F1 on the validation split (n = 256). Final evaluation is performed on the held-out test split (n = 275). All experiments were conducted on an NVIDIA RTX A4000 GPU (16.9 GB VRAM). All training scripts, model checkpoints, and random seeds are available at `https://github.com/misellabmail/strokeid-ner`. Table 3 summarizes the training hyperparameters for each model.

| Hyperparameter | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| Epochs | 10 | 12 | 10 | 10 |
| Learning rate | 1e-5 | 1e-5 | 5e-6 | 1e-5 |
| Batch size | 16 | 16 | 4 (accum. ×2) | 8 |
| Effective batch size | 16 | 16 | 8 | 8 |
| Max sequence length | 128 | 128 | 128 | 128 |
| Optimizer | AdamW | AdamW | AdamW | AdamW |
| Weight decay | 0.01† | 0.01† | 0.01† | 0.01† |
| Warmup ratio | 20% | 20% | 10% | 20% |
| Gradient clipping | 1.0 | 1.0 | 1.0 | 1.0 |
| Precision | FP32 | FP32 | FP16 | FP32 |
| Gradient checkpointing | No | No | Yes | No |

**Table 3.** Training hyperparameters for fine-tuned models. † *Weight decay applied to non-bias, non-LayerNorm parameters only, following Devlin et al.[30] and Loshchilov and Hutter[31]*

XLM-RoBERTa-large requires more conservative training settings due to its 560M parameter count, using a lower learning rate (5e-6 vs. 1e-5), gradient accumulation (effective batch size 8), FP16 mixed precision, and gradient checkpointing. XLM-RoBERTa-large was trained with a fixed random seed; macro F1 = 0.7823 is reported.

### Binary NER Experiment

Following joint multi-entity training, we conduct an additional experiment to investigate whether entity-specific fine-tuning improves recognition performance. We train four binary classifiers using XLM-RoBERTa-large, each specialized for a single entity type. For each binary model, the label space is reduced from 9 BIO tags to 3 tags — O, B-X, and I-X — where X is the target entity type. All other hyperparameters are identical to the joint XLM-RoBERTa configuration described in Section 2.9. Results are presented and analyzed in Sections 3.1 and 3.5. The joint multi-entity XLM-RoBERTa-large results reported in Table 6 are derived from the best single checkpoint (macro F1 = 0.7780), while Table 4 reports the single run result (F1 = 0.7823). The RF F1 of 0.81 in Table 6 similarly reflects single-checkpoint performance and differs from the RF F1 of 0.78 reported in Table 4.

### Evaluation Metrics

We report macro-averaged Precision, Recall, and F1-score across all four entity types using span-level evaluation via seqeval[24]. Macro-averaging treats each entity type with equal weight regardless of frequency, which is appropriate given that all four entity types carry equal clinical significance in stroke triage applications, a rare risk factor mention is no less clinically important than a frequent symptom mention.

To avoid implicitly down-weighting less common entity types, we use macro-averaged F1 instead of micro-averaged F1. SYM (2,643 spans) would contribute disproportionately to RF (839 spans) under micro-averaging. Rather than variations in clinical significance, the observed class frequency differences are a reflection of the nature of patient consultation queries.

Compared to token-level evaluation, span-level evaluation is more stringent and more representative of real-world information extraction utility since it demands precise boundary and label matching.

## Results

Test set performance for each of the five systems on StrokeID-NER (n=275) is shown in Table 4. XLM-RoBERTa-large with a linear classification head outperforms all IndoBERT variants and the GPT-5.4 zero-shot baseline, achieving the best overall macro F1 of 0.7823. The zero-shot LLM baseline is consistently outperformed by all four fine-tuned models, suggesting that domain-specific fine-tuning on StrokeID-NER offers significant advantages over general-purpose language model capabilities alone.

| Model | P | R | F1 | DGN | RF | SYM | TMP |
|---|---|---|---|---|---|---|---|
| GPT-5.4 (zero-shot) | 0.6762 | 0.6659 | 0.6682 | 0.88 | 0.55 | 0.55 | 0.69 |
| IndoBERT-base + linear (M1) | 0.6976 | 0.7818 | 0.7367 | 0.88 | 0.73 | 0.67 | 0.67 |
| IndoBERT-base + CRF (M2) | 0.7017 | 0.7896 | 0.7427 | 0.86 | 0.73 | 0.68 | 0.71 |
| IndoBERT-large + linear (M4) | 0.7264 | 0.8004 | 0.7614 | 0.89 | 0.76 | 0.68 | 0.71 |
| XLM-RoBERTa-large + linear (M3) | 0.7468 | 0.8229 | 0.7823 | 0.87 | 0.78 | 0.73 | 0.75 |

**Table 4.** test set performance on StrokeID-NER (n=275). P = macro precision, R = macro recall, F1 = macro F1. Per-entity scores are F1. Evaluation : span-level seqeval.

Recall consistently outpaces precision in all refined models; this is especially true for XLM-RoBERTa-large (R = 0.8229, P = 0.7468). A tendency toward over-prediction of entity spans is indicated by this systematic pattern. This is the preferred error direction for clinical triage applications because a false positive carries a lower clinical risk than missing a stroke symptom or risk factor.

The surface-form regularity of stroke diagnosis expressions, particularly the term stroke and its variants, which make up 61.1% of all DGN spans, is reflected in DGN's highest F1 among all fine-tuned models (0.86-0.89). Interestingly, IndoBERT-large has the highest DGN F1 of any model (0.89), suggesting that monolingual scale offers a special benefit for identifying Indonesian stroke terminology. SYM has the lowest F1 of all the fine-tuned models (0.67-0.73), which is in line with its extreme surface-form variation reported in Section 2.6 and its 80.0% hapax rate.

The GPT-5.4 zero-shot baseline achieves macro F1 = 0.6682, substantially below all fine-tuned models. On DGN (F1 = 0.88), GPT-5.4 performs similarly to refined models, but significantly lags on RF (F1 = 0.55) and SYM (F1 = 0.55), entity types that call for the recognition of informal Indonesian expressions, Javanese lay terminology, and clinical contextual knowledge that zero-shot prompting cannot consistently capture. A detailed comparative error analysis is given in Section 3.5.

### Effect of CRF Decoding

Comparing M1 (IndoBERT-base + linear, F1 = 0.7367) with M2 (IndoBERT-base + CRF, F1 = 0.7427) reveals that CRF decoding provides a modest overall improvement of 0.006 macro F1 points. A more complex picture is shown by the per-entity breakdown in Table 5. CRF provides the largest benefit for TMP (+0.04), where temporal expressions tend to follow syntactically consistent internal patterns that benefit from structured sequential decoding, and a marginal improvement for SYM (+0.01). In contrast, CRF marginally degrades DGN (−0.02), while RF shows no meaningful change (0.00). We attribute the DGN degradation to the contextually dependent nature of diagnosis recognition: the transition constraints imposed by the CRF layer may penalize valid BIO sequences when diagnosis mentions co-occur with surrounding entity types. The absence of any RF change suggests that risk factor recognition depends primarily on broader clinical context rather than local sequential patterns, which the CRF layer neither helps nor hinders.

The modest overall improvement from CRF (+0.006 points) suggests that IndoBERT's self-attention mechanism already captures much of the sequential dependency information that CRF traditionally provides, consistent with findings in prior BERT-based NER work. Structured decoding benefits entities with consistent internal syntactic structure (TMP, SYM), while having negligible or negative effect on entities whose recognition depends primarily on broader clinical context (DGN, RF).

### Effect of Model Scale and Pretraining Language

Comparing M1 (IndoBERT-base, F1 = 0.7367) and M4 (IndoBERT-large, F1 = 0.7614) isolates the effect of model scale within the monolingual Indonesian pretraining family. Increasing from 124M to 334M parameters yields a consistent improvement of 2.5 macro F1 points across all entity types, with the largest gains for RF (+0.03) and DGN (+0.01).

Comparing M4 (IndoBERT-large, F1 = 0.7614) and M3 (XLM-RoBERTa-large, F1 = 0.7823) isolates the effect of multilingual pretraining while controlling for model scale, as both architectures use 1,024-dimensional hidden states. Multilingual

| Entity | Linear | CRF | Δ | |
|---|---|---|---|---|
| DGN | 0.88 | 0.86 | −0.02 | ↓ |
| RF | 0.73 | 0.73 | 0.00 | - |
| SYM | 0.67 | 0.68 | +0.01 | ↑ |
| TMP | 0.67 | 0.71 | +0.04 | ↑ |
| **Macro** | **0.7367** | **0.7427** | **+0.006** | ↑ |

**Table 5.** Per-entity F1 : IndoBERT-base linear vs CRF (Δ values shown on 0-1 scale)

pretraining provides an additional 2.1 F1 points beyond monolingual scale. This gap is most pronounced for RF (+2 F1 points), and SYM (+5 F1 points), precisely the two entities with the highest hapax rates and the greatest prevalence of informal, code-mixed, and regionally varied expressions. XLM-RoBERTa's pretraining on 100 languages provides richer subword representations for the orthographic variation, informal spelling, and code-mixing characteristic of patient-generated Indonesian health text. These results indicate that multilingual pretraining contributes more to performance than monolingual scale alone for informal clinical NER in Indonesian.

### Binary NER Experiment

Table 6 compares entity-specific binary fine-tuning against joint multi-entity training using XLM-RoBERTa-large. Note that the joint F1 in Table 6 (0.7780) reflects the checkpoint used for this experiment and differs from the main benchmark result in Table 4 (0.7823), as described in Section 2.10. Binary models improve TMP (+0.02 F1) and DGN (+0.01 F1), indicating these entities benefit from focused training without label competition. However, RF degrades substantially in binary mode (−0.08 F1), the largest single-entity change across all experiments.

| Entity | Joint F1 | Binary F1 | Δ | |
|---|---|---|---|---|
| DGN | 0.87 | 0.88 | +0.01 | ↑ |
| TMP | 0.75 | 0.77 | +0.02 | ↑ |
| SYM | 0.72 | 0.72 | ≈0.00 | — |
| RF | 0.81 | 0.73 | −0.08 | ↓ |
| **Macro** | **0.7780** | **0.7726** | **−0.005** | ↓ |

**Table 6.** Joint vs binary NER – XLM-RoBERTa-large (test set) Δ values shown on 0-1 scale

We attribute the RF degradation to the contextual dependency of risk factor identification: the binary RF model cannot observe co-occurring SYM and DGN tokens that provide critical signals for distinguishing clinical risk factors from general health mentions. In the joint model, the presence of a stroke diagnosis provides context that *hipertensi* (hypertension) is a stroke-related risk factor rather than a general health mention a signal unavailable to the binary model. SYM exhibits no significant change (≈0.00), indicating that lexical coverage rather than label competition is its main obstacle. The joint model's overall macro F1 is higher (0.7780 vs. 0.7726 binary), suggesting that this task benefits from cross-entity context.

### Comparative Error Analysis : fine-tuned vs. zero-shot LLM

We looked at span-level prediction overlap to describe the performance difference between the GPT-5.4 zero-shot baseline and the optimal fine-tuned model (M3 XLM-RoBERTa-large) at a finer granularity than aggregate F1 scores. Four possible outcomes were obtained by noting whether each model accurately predicted each gold entity span in the test set: either both models were correct, M3 was correct only, GPT-5.4 was correct only, or both were incorrect. Table 7 provides a summary of these counts.

Out of the 261 gold spans where the two models disagree, M3 correctly identifies 204 spans that GPT-5.4 missed, whereas GPT-5.4 correctly identifies only 57 spans that M3 missed. This indicates a 3.6:1 ratio that favors the improved model. This consistent asymmetric error pattern provides descriptive evidence that the 11.4 F1-point performance gap reflects systematic model behavior rather than random fluctuation.

The per-entity decomposition indicates that this asymmetry is concentrated in specific entity types. For SYM, M3 correctly identifies 125 spans that GPT-5.4 overlooked, while only 27 in the opposite direction (4.6:1 ratio). The asymmetry is even more noticeable for RF: 42 versus 6 (7.0:1 ratio). The two models perform nearly identically for DGN, with M3 winning on 18 spans and GPT-5.4 winning on 15 (1.2:1 ratio). M3 wins on 19 spans for TMP, while GPT-5.4 wins on 9 (2.1:1 ratio). These per-entity

| Outcome | Count | Proportion | Notation |
|---|---|---|---|
| Both models correct | 681 | 62.2% | *n11* |
| M3 correct, GPT-5.4 incorrect | **204** | **18.6%** | *n10* |
| GPT-5.4 correct, M3 incorrect | 57 | 5.2% | *n01* |
| Both models incorrect | 153 | 14.0% | *n00* |
| **Total gold spans** | **1,095** | **100.0%** | |

**Table 7.** Span-level predictions overlap between M3 XLM-RoBERTa-large and GPT-5.4 zero-shot, computed across all gold entity spans in the test set (n=1,095 spans)

ratios show that the value of fine-tuning is concentrated in entities that need to recognize clinical contextual knowledge and informal Indonesian expressions exactly the qualities that separate patient-generated health queries from formal clinical text that general-purpose LLMs are exposed to during pretraining. The surface-form-regular term stroke and its variations, on the other hand, dominate DGN recognition, and GPT-5.4 can handle them just as well as the refined model.

### Error Analysis

Three main error patterns are revealed by a thorough error analysis of XLM-RoBERTa-large test set predictions. First, span boundary errors on SYM predominate; the model starts or extends SYM spans at adjacent non-entity tokens incorrectly, reflecting the extreme lexical diversity of symptom expressions, which range from extended creative descriptions (*terasa seperti menggunakan* high heels *sebelah*) to single-token forms (*lemas*, *pusing*). Second, the model's propensity to incorporate surrounding modifying terms like *penyakit* (disease) or *serangan* (attack) within diagnosis spans is reflected in DGN boundary extension errors. Third, TMP misidentification can cause non-temporal modifiers to be incorrectly categorized as temporal expressions. The top errors are notably devoid of cross-entity label confusion errors; the model hardly ever predicts SYM as DGN or RF as TMP. This indicates that rather than entity type boundaries, which are linguistically distinct, the primary challenge is determining span boundaries.

### Hapax Rate and Entity Difficulty

Lexical variety and model performance are negatively correlated for three of the four entity types shown in Figure 3, which shows the relationship between entity-level hapax rate and XLM-RoBERTa-large F1 score.

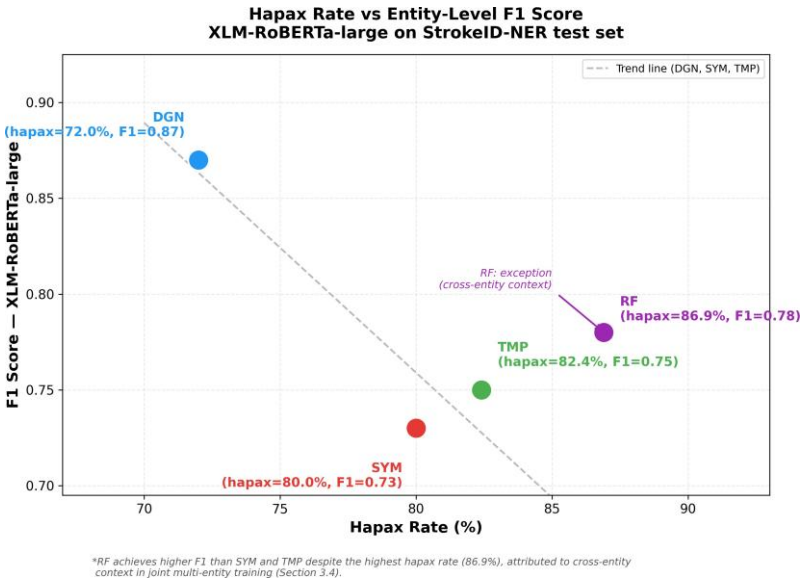



**Figure 3.** Relationship between hapax rate and entity-level F1 score for XLM-RoBERTa-large on the StrokeID-NER test set.

DGN exhibits the lowest hapax rate (72.0%) and the highest F1 score (0.87). The term "stroke" occurs 1,338 times across 2,189 DGN spans (61.1), providing a strong lexical anchor that can be effectively leveraged by both GPT-5.4 and fine-tuned models. SYM shows a contrasting scenario: 1,061 of 1,326 unique symptom expressions occur just once because of its 80%

hapax rate, despite having the highest number of training spans (2,643). While testing, the model encounters creative patient descriptions it hasn't previously encountered, like "*serasa darahnya ngak ngaliiiir*, *drodog*, *cekot-cekot*" for which there's no similar training signal. This indicates that, even with the largest training dataset, SYM achieves the lowest F1 (0.73), implying that lexical coverage is more of a constraint than the amount of training data.

Despite having the highest hapax rate (86.9%), RF achieves F1 = 0.78, which is higher than both TMP (82.4%, F1 = 0.75) and SYM (80.0%, F1 = 0.73). RF is a deviation from the hapax F1 pattern. We credit this to two reasons: (1) RF expressions are clinically limited; patients frequently mention familiar medical issues (diabetes, hypertension, and cholesterol) that the model identifies despite spelling differences; and (2) the cross-entity context in the joint model greatly enhances RF recognition, illustrated by the −0.08 F1 drop in the binary RF test (Section 3.4). These findings indicate that the primary factor driving performance differences in patient-generated clinical NER is lexical diversity, not class frequency or the complexity of entity definitions.

### Cross-Model Error Consistency

We examined the intersection of false negatives, or undetected gold entity spans, across all four refined models to assess if the error patterns indicate flaws specific to the models or if they reflect systematic difficulties at the dataset level. Simultaneously, all four models fail to capture 68.3% of the false negative spans. This significant overlap indicates that the most difficult spans likely represent genuine language challenges rather than being specific blind spots of the model.

The spans that are often missed are hapax SYM expressions. These include unique patient metaphors, Javanese lay terms, and informal descriptions that appear only once in the training set. No fine-tuned model can reliably recognize an expression it hasn't seen in training, no matter the pretraining language, model scale, or decoding strategy. The findings provide a clear insight: improving performance in these spans requires enhancing the training vocabulary. This can be done by adding more annotations, using data augmentation, or applying domain-adaptive pretraining. It's not about changing the model architecture. The 68.3% cross-model overlap indicates a performance ceiling specific to the current dataset, not the NER architecture tested.

## Discussion

The main conclusion of this study is that multilingual pretraining greatly benefits informal patient-generated clinical text. It outperforms monolingual Indonesian pretraining at the same model scale. XLM-RoBERTa-large, fine-tuned on StrokeID-NER, achieves the highest overall performance (macro F1 = 0.7823) across all five systems tested. Three additional findings also shape future research in this area.

Lexical diversity, not entity frequency, mainly affects how hard each entity is. Despite having the most training spans (2,643), SYM has the lowest F1 among fine-tuned models (0.67–0.73) because 80.0% of its 1,326 distinct surface forms only appear once. All four of the refined models share this structural limitation: 68.3% of false negative spans are missed by all four at the same time, suggesting that the training data, rather than the model architecture imposed a performance ceiling. Instead of making architectural changes, increasing training vocabulary through more annotation or data augmentation is necessary to improve performance on these spans.

Second, the value of fine-tuning over zero-shot prompting is entity-specific rather than uniform. Fine-tuned XLM-RoBERTa correctly identified 204 gold spans that GPT-5.4 missed against only 57 in the opposite direction (3.6:1 ratio), with the advantage concentrated in SYM (4.6:1) and RF (7.0:1) entities requiring recognition of informal Indonesian expressions and clinical contextual knowledge. For DGN, the two approaches perform equivalently, reflecting the dominance of the surface-form-regular term stroke (61.1% of all DGN spans) which general-purpose LLMs recognize as effectively as fine-tuned models. This finding is consistent with recent evidence that fine-tuned BERT-based models outperform zero-shot LLM approaches for clinical NLP tasks[5,32] and extends this observation to informal patient-generated text in a low-resource language.

Third, multilingual pretraining contributes more to performance than monolingual scale alone. The comparison between M4 (IndoBERT-large, F1 = 0.7614) and M3 (XLM-RoBERTa-large, F1 = 0.7823), both using 1,024-dimensional architectures, yields a 2.1 F1-point advantage concentrated in SYM (+5 points) and RF (+2 points) precisely the entities with the highest hapax rates and the greatest prevalence of informal, code-mixed, and regionally varied expressions. The exception is DGN, where IndoBERT-large achieves the highest F1 of any model (0.89), consistent with the surface-form regularity of stroke terminology in Indonesian text. These findings suggest that the choice of pretraining language interacts with entity characteristics: multilingual pretraining is preferable for entities with high lexical diversity, while monolingual scale may retain marginal advantages for entities dominated by high-frequency regular surface forms.

StrokeID-NER addresses a domain not covered by prior Indonesian medical NER work. Kusumawardani and Kusumawati[11] developed a general health consultation NER system across multiple topics on a mixed-register corpus combining patient and physician utterances, which is not directly comparable to our patient-only stroke-specific setting. Ignasius et al.[12] constructed the DetikHealth corpus from formal health-news articles, reporting a best macro F1 of 0.3297 with BiLSTM-CRF. The substantially higher performance in the present study (macro F1 = 0.7823) primarily reflects the transition from BiLSTM to

transformer-based architectures: our M1 IndoBERT-base alone achieves macro F1 = 0.7367, confirming the well-established advantage of direct transformer fine-tuning over BiLSTM-based approaches[5]. The key distinguishing contribution of StrokeID-NER is the exclusive focus on patient-generated informal text, where hapax rates of 72.0–86.9% represent a challenge that corpora based on formal registers do not capture.

### Limitations

Several limitations of this study should be acknowledged. First, the test set size (n = 275 queries) is modest; observed F1 differences between adjacent fine-tuned models ($\Delta$F1 = 0.006–0.021) should be interpreted with appropriate caution. Second, the dataset was collected from a single platform (Alodokter.com) during a single collection period. While Alodokter is the largest Indonesian health consultation platform, generalization to other platforms, clinical contexts, or regional language distributions cannot be assumed without additional evaluation. Third, annotation was performed by a single primary annotator, though the inter-annotator agreement assessment ($\kappa$ = 0.857) on an independently annotated stratified sample confirms the reliability of the annotation process. Fourth, XLM-RoBERTa-large showed continued improvement through epoch 10 without convergence, suggesting that additional training steps beyond our 10-epoch budget may yield further performance gains; this was not pursued due to computational resource constraints. Finally, this study evaluates NER performance as a linguistic task; clinical deployment utility including integration into triage workflows and evaluation by clinical practitioners requires prospective validation beyond the scope of the present work.

### Conclusions

This study introduces StrokeID-NER, the first publicly released stroke-specific NER dataset for informal Indonesian patient-generated health queries, comprising 1,742 queries with 6,765 annotated spans across four clinically grounded entity types. Benchmarking five NER systems reveals that fine-tuned XLM-RoBERTa-large achieves the best performance (macro F1 = 0.7823), with multilingual pretraining providing a consistent advantage over monolingual Indonesian pretraining at equivalent model scale for entities characterized by high lexical diversity and informal register. Fine-tuning on StrokeID-NER substantially outperforms zero-shot GPT-5.4 prompting, with a 3.6:1 ratio of unique correct predictions in favor of the fine-tuned model, confirming that domain-specific annotation remains essential for clinical NLP in low-resource informal settings even when state-of-the-art general-purpose LLMs are available. Linguistic analysis reveals that hapax rate, rather than entity frequency, is the primary driver of per-entity difficulty, and that 68.3% of false negative spans represent a shared structural ceiling across all four fine-tuned architectures. These findings show that improving performance is best achieved by expanding training vocabulary. This can be done through more annotation or data augmentation, not just by changing the architecture. StrokeID-NER, the annotation guidelines, and all trained models are publicly available to support future Indonesian clinical NLP research.

## References


1. Venketasubramanian, N., Yudiarto, F. & Tugasworo, D. Stroke burden and stroke services in indonesia. *Cerebrovasc. Dis. Extra* **12**, 53–57, DOI: 10.1159/000524161 (2022).
2. Harmadha, W. *et al.* Cardiovascular disease mortality in Indonesia: analysis of Global Burden of Disease 2000–2019. *PLOS ONE* DOI: 10.1371/journal.pone.0294128 (2023).
3. Zainuddin, A., Kadir, R., Kuswanto, H. *et al.* The major risk factor of stroke across Indonesia; a nationwide geospatial analysis of universal health coverage program. *Arch. Public Heal.* **83**, 169, DOI: 10.1186/s13690-025-01613-4 (2025).
4. Sezgin, E., Hussain, S., Rust, S. & Huang, Y. Extracting medical information from free-text and unstructured patient-generated health data using NLP: feasibility study with real-world data. *JMIR Form. Res.* **7**, e43014, DOI: 10.2196/43014 (2023).
5. Gonzalez Hernandez, F., Nguyen, Q., Smith, V. *et al.* Named entity recognition of pharmacokinetic parameters in the scientific literature. *Sci. Reports* **14**, 23485, DOI: 10.1038/s41598-024-73338-3 (2024).
6. Khairunnisa, S. *et al.* Towards a standardized dataset on Indonesian named entity recognition. In *Proceedings of the 1st AACL Student Research Workshop*, DOI: 10.18653/v1/2020.aacl-srw.10 (2020).
7. Koto, F., Rahimi, A., Lau, J. H. & Baldwin, T. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770, DOI: 10.18653/v1/2020.coling-main.66 (2020).
8. Aji, A. *et al.* One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 7226–7249, DOI: 10.18653/v1/2022.acl-long.500 (2022).

9. Winata, G. *et al.* NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, DOI: 10.18653/v1/2023.eacl-main.57 (2023).
10. Zainuddin, Z., Mudassir & Tahir, Z. Entity extraction in Indonesian online news using NER with hybrid method Transformer, Word2Vec, Attention and Bi-LSTM. *JOIV: Int. J. on Informatics Vis.* **9**, 964–973, DOI: 10.62527/joiv.9.3.2902 (2025).
11. Kusumawardani, R. & Kusumawati, K. Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-LSTM-CRF algorithm. *Procedia Comput. Sci.* **245**, 1146–1156, DOI: 10.1016/j.procs.2024.10.344 (2024).
12. Ignasius, D., Dewi, I. N., Norman, M. B. C. & Sani, R. R. Medical named entity recognition from Indonesian health-news using BiLSTM-CRF with static and contextual embeddings. *J. Appl. Informatics Comput.* **9**, 2974–2985, DOI: 10.30871/jaic.v9i6.11574 (2025).
13. Powers, W. *et al.* 2018 guidelines for the early management of patients with acute ischemic stroke. *Stroke* **49**, e46–e110, DOI: 10.1161/STR.0000000000000158 (2018).
14. Kementerian Kesehatan Republik Indonesia/PERDOSSI. Panduan nasional pelayanan kedokteran (PNPK) stroke. Tech. Rep., Kemenkes RI, Jakarta (2019). No. HK.01.07/MENKES/394/2019.
15. World Health Organization. ICD-10 version 2019. https://icd.who.int/browse10/2019/en (2019).
16. Kementerian Kesehatan Republik Indonesia. Laporan nasional Riskesdas 2018. Tech. Rep., Badan Litbangkes, Jakarta (2018).
17. Abdillah, A. F. indonesia-medical-qna (revision c8c08cd) [dataset]. Hugging Face, DOI: 10.57967/hf/8356 (2026).
18. Aroor, S., Singh, R. & Goldstein, L. BE-FAST (Balance, Eyes, Face, Arm, Speech, Time): reducing the proportion of strokes missed using the FAST mnemonic. *Stroke* **48**, 479–481, DOI: 10.1161/STROKEAHA.116.015169 (2017).
19. Tkachenko, M. *et al.* Label Studio: data labeling software. https://labelstud.io (2020).
20. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46, DOI: 10.1177/001316446002000104 (1960).
21. Landis, J. & Koch, G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174, DOI: 10.2307/2529310 (1977).
22. Uzuner, Ö., South, B., Shen, S. & DuVall, S. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Informatics Assoc.* **18**, 552–556, DOI: 10.1136/amiajnl-2011-000203 (2011).
23. Tjong Kim Sang, E. F. & De Meulder, F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147 (2003).
24. Nakayama, H. seqeval: a Python framework for sequence labeling evaluation. https://github.com/chakki-works/seqeval (2018).
25. Wilie, B. *et al.* IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857, DOI: 10.18653/v1/2020.aacl-main.85 (2020).
26. Lafferty, J., McCallum, A. & Pereira, F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289 (2001).
27. Wirawan, C. xlm-roberta-large-indonesian-NER [model]. HuggingFace. https://huggingface.co/cahya/xlm-roberta-large-indonesian-NER (2020).
28. Conneau, A. *et al.* Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451, DOI: 10.18653/v1/2020.acl-main.747 (2020).
29. OpenAI. GPT-5.4. https://openai.com (2026).
30. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186, DOI: 10.18653/v1/N19-1423 (2019).
31. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv:1711.05101* DOI: 10.48550/arXiv.1711.05101 (2019).

**32.** Diaz Ochoa, J. *et al.* Optimized BERT-based NLP outperforms zero-shot methods for automated symptom detection in clinical practice. *Front. Digit. Heal.* **7**, 1623922, DOI: 10.3389/fdgth.2025.1623922 (2025).

## Ethics Statement

This study used the publicly available Indonesia-medical-qna dataset[17] from alodokter.com. The data has no personal identifiable information. An ethical review was waived since the study only uses public, fully anonymized, third-party data. Patient queries were submitted voluntarily to a public platform. No additional consent was needed because there was no direct patient contact.

## Data Availability

The StrokeID-NER models and labeled dataset are now open source. You can find them at `https://doi.org/10.5281/zenodo.20305007` and `https://github.com/misellabmail/strokeid-ner`. They are under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The source data comes from a publicly available HuggingFace dataset[17]. It is intended only for academic research, following the platform's terms of use. Note that the source dataset does not have an explicit license.

## Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) - Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00436773).

## Funding Declaration

This research was funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) - Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00436773).

## Author contributions statement

M.C.S conducted the experiments and wrote articles. M.Y and M.A.I.M contributed to methodology development. H.C.K supervised the research and provided clinical guidance. All authors reviewed the manuscript.

## Additional information

**Competing Interests:** The authors declare no competing interest.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryNote.docx