VIMED-STRUCT: STRUCTURED EMR GENERATION FROM VIETNAMESE CLINICAL SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical documentation requires physicians to repeatedly convert information conveyed during patient encounters into structured electronic medical record (EMR) fields, creating a substantial workload. Automating this process from speech is promising, yet progress is limited by the scarcity of paired clinical audio and structured EMR annotations. This bottleneck stems not only from privacy and practical constraints on clinical audio recording, but also from the high cost of structured data annotation, particularly for low-resource languages. To address this gap, we introduce a framework for constructing aligned clinical audio - structured EMR datasets from hospital records. By deriving structured targets directly from existing EMR data and separating them from speech realization, the framework removes the need for costly manual structured annotation while enabling scalable diversification of speakers, speaking styles, and acoustic conditions. Using this framework, we build **ViMed-Struct**, to the best of our knowledge the first large-scale Vietnamese speech-to-structured-EMR dataset, comprising 2,572 clinical cases and 2,892 recordings totaling 47.99 hours, collected from 58 speakers under matched quiet and noisy conditions. We further establish a benchmark with a field-aware evaluation metric that measures both correct field selection and extracted value quality, providing a more faithful evaluation than conventional text-similarity metrics alone. Finally, we propose an ASR - Information Extraction framework for speech-to-structured-EMR prediction and compare it with representative existing approaches. Experiments on ViMed-Struct show that our method achieves up to 94% macro-averaged field-presence F1, demonstrating strong performance in structured extraction across diverse backbone models.

1 INTRODUCTION

Background. Clinical documentation remains an essential but labor-intensive part of healthcare. Much of the information needed for patient records is naturally expressed in free form speech, for example when clinicians summarize a patient's history, examination findings, investigations, diagnoses, or treatment plans (Shi et al., 2024; Delbrouck et al., 2025). In contrast, hospital information systems require these observations to be entered into predefined structured fields. Speech recognition alone therefore does not solve the documentation problem. After transcription, the system must still identify relevant clinical information, map it to the appropriate schema fields, and populate the corresponding values(Krishna et al., 2021). Recent advances in automatic speech recognition (ASR) and large language models (LLMs) have made this workflow increasingly tractable(Shih et al., 2026). Prior studies have explored spoken nursing handovers, structured nursing documentation, observation extraction, medical order generation, and voice-driven EHR form filling (Suominen et al., 2015; Corbeil et al., 2025a;b; Su et al., 2026; Tanhaei, 2026). A key challenge is adapting such systems to clinical domains where paired speech-to-structure data are scarce (Le-Duc, 2024). A fully end-to-end speech-to-structure system requires supervision linking spoken input directly to structured targets, which can be costly to collect for each new specialty, institution, or documentation schema. In contrast, decomposing the task into automatic speech recognition (ASR) and information extraction (IE) allows the two components to be adapted independently, with ASR leveraging schema independent speech data and IE using text or structured data from the target domain. At the same time, it raises an important question: how much of the remaining difficulty in clinical speech-to-structure prediction comes from speech recognition itself, and how much arises from mapping the recovered clinical information to the correct structured fields?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Research Gap. Current clinical speech-to-structure approaches leave two gaps. First, paired clinical speech and structured EMR data remain difficult to obtain at scale. Collecting clinical audio is constrained by privacy and recording requirements, while structured targets require substantial annotation, making paired data particularly scarce in low-resource settings(Liu & Liu, 2026). Existing benchmarks are therefore largely centered on specific documentation settings, such as nursing observations, medical orders, handover categories, or individual EHR forms, rather than structured patient records grounded in routine hospital data and spanning multiple stages of care. Second, conventional evaluation can obscure important dimensions of clinical model performance that are not captured by aggregate end-to-end metrics (Arias-Duart et al., 2025; Jing et al., 2026; Sepehri et al., 2025). Metrics such as ROUGE-L or BERTScore primarily assess value similarity and, when aggregated across fields, do not explicitly distinguish field selection errors from value extraction errors(Abacha et al., 2023; Chen et al., 2023). For sparse clinical schemas, reliable evaluation therefore requires measuring field selection separately from the quality of the extracted values.

Our Contribution. To address these limitations, we introduce **ViMed-Struct**, a clinical speech-to-structured-EMR benchmark constructed from existing hospital records. Our primary contributions are summarized as follows:

- We introduce a record-grounded framework for constructing aligned clinical speech-structured EMR datasets from existing hospital records. The resulting resource provides a comprehensive testbed for evaluating the full speech-to-structured-EMR task and a practical foundation for developing clinical documentation systems in low resource settings. Using this framework, we construct **ViMed-Struct**, comprising 2,572 text structure cases, including 1,475 audio linked cases represented by 2,892 recordings totaling 47.99 hours from 58 speakers under matched quiet and noisy conditions.
- We establish a comprehensive benchmark spanning direct speech-to-structure models, ASR and LLM cascades, and reference transcript extraction. By separating field selection from value extraction, our analysis reveals systematic field-assignment failures that remain hidden under conventional content similarity evaluation.
- Finally, to address the observed field-level imbalance, we propose Macro Field Loss, a field aware fine-tuning objective that rebalances the learning signal across schema fields. This objective improves structured extraction across multiple LLM backbones, achieving up to 94% macro-averaged field-presence F1 while maintaining comparable value extraction quality.

2 Related Work

2.1 Clinical speech-to-structure datasets and benchmarks

Clinical speech processing has increasingly moved beyond transcription toward structured documentation. Early work on the NICTA Synthetic Nursing Handover corpus combined automatic speech recognition (ASR) with information extraction to map spoken nursing handovers into predefined handover categories, while CLEF eHealth subsequently extended this setting with shared task annotations and evaluation (Suominen et al., 2015; 2016). More recent studies consider documentation workflows closer to clinical practice (Kernberg et al. (2024)). Voice-to-DART converts nursing speech into structured Data, Action, Response, and Teaching records for hospital information systems (Su et al., 2026), while BioArc maps clinician dictations into heterogeneous EHR forms containing free text, categorical, numerical, and nested fields (Tanhaei, 2026). These resources establish the feasibility of speech-driven structured documentation across different clinical settings. However, they differ considerably in the scope of information being structured and, importantly, in how their structured targets are grounded in routine clinical records Corbeil et al. (2025a) ViMed-Struct is constructed from routine electronic clinical records, from which a canonical structured target is derived for each case before narrative and speech realization. This design preserves case-level grounding across the source clinical record, structured target, corresponding narrative, and recorded speech. The resulting representation covers admission information, medical history, examination, investigations, diagnoses, treatment, and clinical outcome. Table 1 summarizes related datasets.

Table 1: Comparison of clinical datasets and benchmarks for speech and transcript based structure extraction. Clinical-record-grounded settings use patient-level structured clinical records as targets.

Dataset	Lang.	Audio	Transcript	Schema filling	Mixed-type fields	Field-presence eval.	Clinical-record-grounded	Paired quiet/noisy	Size
NICTA / CLEF (Suominen et al., 2015; 2016)	EN	✓	✓	✓	✗	✗	✗	✗	101 handover records
SYNUR (Corbeil et al., 2025a)	EN	✗	✓	✓	✓	✗	✗	✗	223 dictations
SIMORD / MEDIQA-OE (Corbeil et al., 2025b)	EN	✗	✓	✓	✓	✗	✗	✗	264 conversations
Voice-to-DART (Su et al., 2026)	ZH, EN	✓	✓	✓	✗	✓	✗	✗	327 evaluation samples
BioArc (Tanhaei, 2026)	FA, EN	✓	✓	✓	✓	✓	✗	✗	1,000 recordings; ∼15 h
ViMed-Struct	**VI**	✓	✓	✓	✓	✓	✓	✓	**2,572 cases; 1,475 with audio; 47.99 h**

2.2 Structured clinical information extraction and reliability

Structured clinical extraction has been studied from transcripts and narratives. SYNUR maps nurse dictations to typed clinical observations and hospital specific flowsheet schemas (Corbeil et al., 2025a), while SIMORD and MEDIQA-OE extract medical orders and associated attributes from doctor-patient conversations (Corbeil et al., 2025a;b). These benchmarks support observation and order level structuring but remain tied to their respective documentation settings. ViMed-Struct instead considers patient-level structured documentation, where models must recover clinical content and place it within a heterogeneous and sparsely populated schema. This broader setting makes field assignment a central reliability challenge. Reliable structured extraction therefore requires not only producing clinically plausible values, but also determining which schema fields are supported by the available evidence and assigning extracted information to the appropriate location. MIDAS_SYNUR examines absence detection and cross field interference through prompting each field (Sriram & Sahoo, 2026), while SemAnTICA separates candidate concept selection from value extraction through schema retrieval, evidence filtering, and gated extraction (Hwang et al., 2026). Related speech-to-form and source-grounded generation studies report missed, incorrectly populated, unsupported, or misassigned information (Su et al., 2026; Tanhaei, 2026; Lee et al., 2026), showing that accurate content recovery does not necessarily imply correct field placement. Accordingly, our benchmark evaluates field selection and value extraction as complementary aspects of structured prediction.

3 ViMed-Struct

3.1 Overview of ViMed-Struct

ViMed-Struct is designed to construct data for Vietnamese clinical speech-to-structured-EMR prediction using de-identified patient records collected from collaborating hospitals. A straightforward approach is to record clinical examinations or handovers and manually map each recording to a structured record. However, this approach may expose sensitive patient information and requires substantial expert annotation. To address these limitations, we propose a novel EMR-grounded data construction framework named ViMed-Struct. Specifically, ViMed-Struct uses de-identified clinical records as a reference source and constructs the dataset through three stages: (1) clinical record normalization, (2) clinical script generation and verification, and (3) speech recording and alignment. This approach offers three advantages: (i) identifiable information is removed before speech collection; (ii) existing structured records directly provide the ground truth, avoiding the need for manual annotation of audio recordings; and (iii) the framework enables systematic data diversification through the use of different speakers and controlled acoustic and transcription variations. These three stages correspond to steps 1,2 and 3 in the upper part of Figure 1.

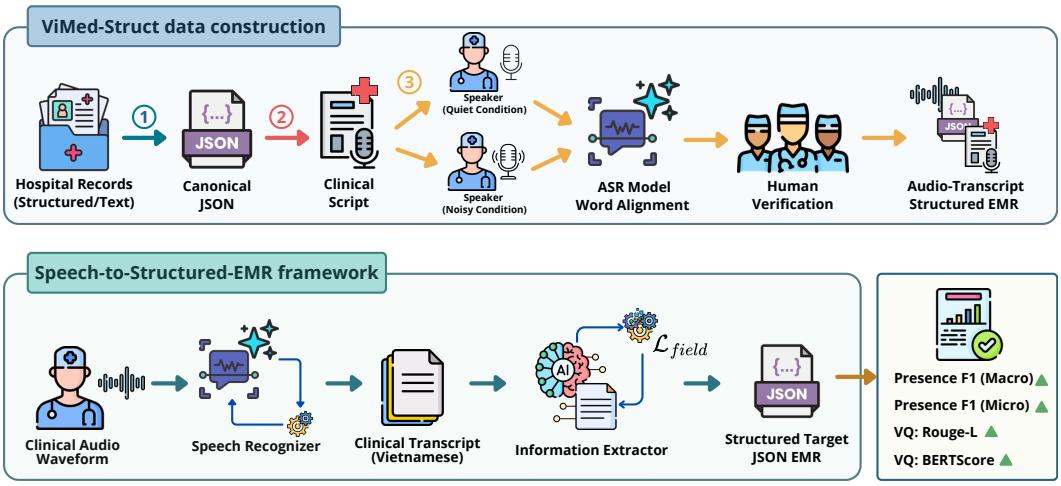

Figure 1: Overview of the ViMed-Struct data construction and the Speech-to-Structured EMR framework. The data construction flow consists of (1) clinical record normalization, (2) clinical script generation and verification, and (3) speech recording and alignment.

3.2 Dataset Construction

1) Clinical record normalization. Patient records are collected from two collaborating hospitals and randomly sampled from the available cases without restriction to a specific clinical department or disease group. Because hospitals may use different record formats, we identify fields with the same clinical meaning across source records and organize them into a shared canonical representation, as further described in Appendix A. This process only establishes the correspondence between source fields and schema fields, without modifying or inferring any clinical content. Missing fields are represented as `null`, while temporally dependent notes are ordered according to their recorded timestamps. The resulting mapped record is used as the canonical structured representation for each case.

2) Clinical script generation and verification. Each canonical structured record is converted into an initial clinical narrative using predefined rules that determine which information is included, how it is ordered, and how structured fields are verbalized (see Appendix B). The script contains only information supported by the source record and does not infer unrecorded clinical events or outcomes. Clinical information is organized according to the progression of a case rather than the order of fields in the schema. Each narrative is constrained to approximately 250-2,000 characters and excludes identifiable information and storage-specific terminology. To enhance data diversity and realism, the initial drafts are edited by merging sentences, reordering phrases and sentences, and adding conversational expressions while preserving the underlying clinical content. Each edited narrative is then checked against its corresponding canonical structured record. Verification focuses on ensuring that clinically important information in the structured record is adequately covered, narrative statements are consistent with the corresponding structured values, no unsupported clinical information is introduced, and the original clinical meaning is preserved after rephrasing and reordering. Narratives that do not satisfy these criteria are revised before being finalized as verified clinical scripts and used as reference scripts for speech recording.

3) Speech recording and alignment. Each verified clinical script is read aloud and recorded by an experienced healthcare worker under a controlled reading protocol. Speakers are instructed to preserve the clinical content, use a natural speaking rate and intonation, avoid adding clinical information or filler content, and verbalize medical abbreviations, numerical values, and measurement units according to common Vietnamese clinical conventions. Recordings are collected from 58 speakers under both quiet and noisy conditions to enhance acoustic diversity; for cases recorded in both conditions, the same reference script is retained to avoid changes in clinical content. The recordings are then aligned with the reference scripts at the word level using PhoWhisper adapted

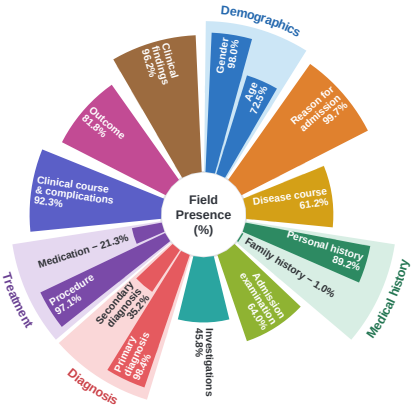

Figure 2: **Field presence across the clinical schema.** Presence denotes the percentage of cases with a non-null, non-empty value for each field.

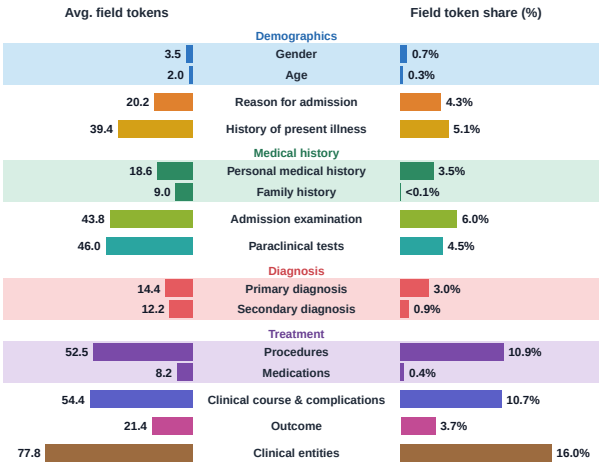

Figure 3: **Field-level token statistics.** Average field tokens are computed over populated occurrences of each field, while field token share denotes the fraction of all structured-target tokens associated with that field.

to Vietnamese medical speech. The resulting timestamps are used for segmentation, after which the audio segments and transcripts are manually reviewed and corrected.

Each case contains a verified clinical script paired with its canonical structured EMR. For cases with recorded speech, we additionally provide audio segments and their corresponding transcripts. Each audio–transcript pair carries a case identifier linking it to the corresponding script-EMR pair.

3.3 Dataset Statistics and Splits

ViMed-Struct contains 2,572 unique clinical cases organized into two case disjoint configurations. **clinical-full** contains 1,475 cases with structured targets, reference narratives, and 2,892 audio recordings totaling 47.99 hours, while **clinical-text** contains an additional 1,097 text structure cases without paired audio, providing a complementary resource for text-based structured extraction. All experiments in this work use **clinical-full**. Table 2 summarizes the corpus and split statistics.

We split **clinical-full** at the case level to prevent leakage between different representations of the same clinical case. Clinical field coverage is computed over the 1,475 cases in **clinical-full**. Figure 2 shows substantial variation in information availability across the schema. The resulting heterogeneous null/non-null patterns make field presence an important component of the prediction problem.

Table 2: Statistics and case partitioning of ViMed-Struct.

Config.	Split	Cases	Rec.	Segments	Dur. (h)
clin-full	Overall	1,475	2,892	50,037	47.99
	Train	1,190	2,325	40,390	38.79
	Validation	142	284	4,853	4.61
	Test	143	283	4,794	4.59
clin-text	Overall	1,097	–	–	–
Total		**2,572**	**2,892**	**50,037**	**47.99**

3.4 Field-Aware Evaluation Protocol

Conventional NLP metrics such as ROUGE-L and BERTScore primarily measure content similarity, but do not fully capture the correctness of structured outputs. A model may preserve clinical content while assigning it to the wrong field, omitting a populated field, or producing content for a field that should be empty, as illustrated in Figure 4. We therefore introduce two complementary metrics: Presence F1 measures *where* information is assigned, whereas Value Quality measures *what* information is extracted.

Presence F1. Presence F1 measures whether the model correctly predicts the presence or absence of information in each field. Ground truth and predicted values are converted into binary null/non-

null decisions for each field. We define its macro- and micro-averaged variants as

$$PF1_{\text{macro}} = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} F1_f, \qquad PF1_{\text{micro}} = \frac{2 \sum_f TP_f}{2 \sum_f TP_f + \sum_f FP_f + \sum_f FN_f}.$$

Here, $F1_f$, TP_f, FP_f, and FN_f denote the F1 score and true positive, false positive, and false negative counts for field f.

Value Quality. Value Quality (VQ) measures the quality of the predicted content for populated free text fields. For each free text field, we evaluate only samples with a non-null ground truth value using ROUGE-L and BERTScore. For $m \in \{\text{ROUGE-L}, \text{BERTScore}\}$, VQ is defined as

$$\text{VQ}^{(m)} = \frac{1}{|\mathcal{F}_{\text{text}}|} \sum_{f \in \mathcal{F}_{\text{text}}} \frac{1}{|\mathcal{S}_f|} \sum_{i \in \mathcal{S}_f} m(p_{i,f}, g_{i,f}), \qquad \mathcal{S}_f = \{i : g_{i,f} \neq \text{null}\}.$$

A null or empty prediction for a populated ground truth value is treated as an empty string. We report $\text{VQ}_{\text{ROUGE-L}}$ and $\text{VQ}_{\text{BERTScore}}$.

4 Benchmarking Clinical Speech-to-Structure Systems

4.1 Experimental Setup

Benchmarked systems. We benchmark two speech-to-structure paradigms: Direct Speech-to-Structure and Cascaded ASR-to-Structure.

In the Direct Speech-to-Structure setting, language models map clinical audio directly to structured outputs; we evaluate Qwen2.5-Omni-7B (Xu et al., 2025) and the commercial model GPT-Audio-1.5. In the Cascaded ASR-to-Structure setting, an ASR system first produces a transcript, which is subsequently processed by a text based structured extractor. We evaluate four recognizers: PhoWhisper-medium, PhoWhisper-large (Le et al., 2024), LoRA adapted PhoWhisper-medium, and Whisper-large-v3 (Radford et al., 2023), paired with Qwen2.5-7B, MedGemma-1.5-4B (Sellergren et al., 2025), Qwen3.5-9B, and Gemma-4-12B (Team et al., 2024). Each extractor is evaluated with zero-shot, few-shot, and LoRA fine-tuning, yielding $4 \times 4 \times 3$ cascaded configurations under quiet and noisy speech. We additionally evaluate the extractors on reference transcripts to isolate structured-extraction errors from ASR errors.

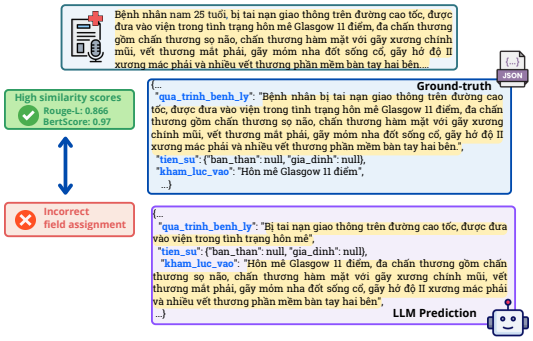

Figure 4: Representative structured-output errors motivating field-aware evaluation.

Research Questions. Our experiments are designed to answer the following research questions:

RQ1: End-to-End Speech-to-Structure Performance. How well do current direct speech models and ASR and LLM cascades transform clinical speech into structured patient records under different acoustic conditions and adaptation settings?

RQ2: Error Decomposition. To what extent are end-to-end errors attributable to speech recognition, and which errors persist when structured extraction is performed from reference transcripts?

RQ3: Field-Level Heterogeneity. What field-level heterogeneity remains in structured extraction once ASR errors are removed, and how is it related to imbalance in the target schema?

Implementation Details. We conduct all experiments on **clinical-full**. PhoWhisper-medium is adapted using LoRA ($r = 16$, $\alpha = 32$, dropout $= 0.05$) for up to 5 epochs with a learning rate of 2×10^{-4}. For the text extractors, we apply LoRA with the same rank, scaling, and dropout to the attention and MLP projections, training for up to 5 epochs with AdamW and an effective batch size of 16. We select the checkpoint with the lowest validation loss for evaluation. Few-shot prompting uses $k = 5$ demonstrations, and inference uses greedy, JSON-schema constrained decoding.

4.2 Results and Analysis

4.2.1 End-to-End Speech-to-Structure Performance

Table 3 shows that end-to-end speech-to-structure performance depends substantially on model architecture, acoustic conditions, and adaptation strategy. Among direct models, GPT-Audio-1.5 achieves a macro Presence F1 of 0.917/0.918 on quiet/noisy speech, indicating that its field prediction is largely robust to noise. The strongest cascaded systems can approach this level: PhoWhisper-medium (LoRA) + Gemma (LoRA) reaches a macro Presence F1 of 0.914 on quiet speech, while achieving a higher VQ-ROUGE-L score than GPT-Audio-1.5 (0.725 vs 0.663).

Table 3: End-to-end speech-to-structure performance under quiet and noisy clinical speech conditions. Bold and underline indicate the best and second-best results, respectively.

ASR	LLM	Set.	PF1-ma		PF1-mi		VQ-R		VQ-B	
			Q	N	Q	N	Q	N	Q	N
Pho Whisper medium	Gemma	ZS	0.873	0.851	0.893	0.876	0.568	0.501	0.730	0.685
		LoRA	**0.896**	**0.897**	**0.906**	**0.909**	**0.652**	**0.596**	**0.821**	**0.811**
		FS	0.866	0.848	0.882	0.866	0.577	0.498	0.743	0.701
	Med Gemma	ZS	0.874	0.858	0.890	0.885	0.565	0.508	0.746	0.717
		LoRA	0.830	0.823	0.844	0.841	0.483	0.413	0.699	0.678
		FS	0.355	0.382	0.374	0.403	0.108	0.098	0.176	0.188
	Qwen2.5	ZS	0.790	0.801	0.810	0.825	0.479	0.446	0.618	0.625
		LoRA	0.748	0.754	0.764	0.775	0.423	0.389	0.580	0.577
		FS	0.632	0.648	0.662	0.680	0.254	0.247	0.389	0.407
	Qwen3.5	ZS	0.817	0.804	0.860	0.850	0.455	0.403	0.655	0.633
		LoRA	0.893	0.869	**0.910**	0.898	0.611	0.534	0.798	0.764
		FS	0.825	0.804	0.865	0.846	0.459	0.406	0.663	0.632
Pho Whisper large	Gemma	ZS	0.882	0.857	0.902	0.883	0.580	0.511	0.741	0.703
		LoRA	**0.894**	**0.898**	**0.905**	**0.909**	**0.650**	**0.592**	**0.823**	**0.810**
		FS	0.865	0.855	0.880	0.872	0.569	0.496	0.741	0.712
	Med Gemma	ZS	0.875	0.862	0.894	0.888	0.577	0.508	0.753	0.724
		LoRA	0.826	0.859	0.843	0.872	0.481	0.446	0.698	0.725
		FS	0.398	0.440	0.418	0.470	0.120	0.115	0.200	0.223
	Qwen2.5	ZS	0.797	0.790	0.821	0.813	0.496	0.438	0.633	0.612
		LoRA	0.751	0.749	0.775	0.776	0.423	0.388	0.584	0.569
		FS	0.624	0.640	0.664	0.670	0.257	0.237	0.384	0.389
	Qwen3.5	ZS	0.821	0.818	0.859	0.855	0.459	0.415	0.660	0.641
		LoRA	0.889	0.881	0.906	**0.905**	0.604	0.534	0.799	0.774
		FS	0.826	0.813	0.864	0.854	0.466	0.410	0.666	0.636
Pho Whisper medium +LoRA	Gemma	ZS	0.906	0.887	0.923	0.906	0.654	0.671	0.769	0.793
		LoRA	**0.914**	**0.897**	**0.909**	**0.925**	**0.725**	**0.715**	**0.842**	**0.861**
		FS	0.888	0.882	0.897	0.905	0.653	0.636	0.784	0.791
	Med Gemma	ZS	**0.913**	0.892	**0.910**	**0.931**	**0.686**	**0.680**	0.790	0.814
		LoRA	0.884	0.895	0.894	0.908	0.668	0.643	0.836	0.842
		FS	0.596	0.600	0.632	0.633	0.274	0.262	0.383	0.382
	Qwen2.5	ZS	0.769	0.841	0.790	0.866	0.543	0.599	0.614	0.704
		LoRA	0.770	0.785	0.786	0.805	0.505	0.503	0.618	0.639
		FS	0.640	0.671	0.672	0.712	0.332	0.343	0.421	0.451
	Qwen3.5	ZS	0.844	0.846	0.885	0.898	0.560	0.543	0.715	0.727
		LoRA	0.885	0.889	0.909	0.922	0.677	0.659	0.808	0.819
		FS	0.841	0.847	0.884	0.895	0.559	0.541	0.713	0.723
Whisper large v3	Gemma	ZS	0.784	0.762	0.836	0.801	0.386	0.373	0.570	0.551
		LoRA	**0.840**	**0.833**	**0.858**	**0.856**	**0.459**	**0.434**	**0.665**	0.657
		FS	0.773	0.752	0.801	0.771	0.373	0.323	0.551	0.517
	Med Gemma	ZS	0.807	0.780	0.847	0.821	0.401	0.349	0.608	0.571
		LoRA	**0.808**	**0.780**	**0.847**	**0.821**	**0.401**	**0.349**	**0.608**	**0.571**
		FS	0.508	0.503	0.539	0.534	0.136	0.118	0.276	0.266
	Qwen2.5	ZS	0.746	0.713	0.794	0.766	0.356	0.305	0.543	0.509
		LoRA	0.637	0.656	0.668	0.689	0.248	0.275	0.405	0.431
		FS	0.555	0.548	0.606	0.591	0.183	0.162	0.296	0.287
	Qwen3.5	ZS	0.736	0.696	0.778	0.735	0.312	0.250	0.503	0.448
		LoRA	**0.842**	0.821	**0.871**	0.856	0.424	0.398	**0.681**	0.653
		FS	0.744	0.707	0.784	0.742	0.320	0.260	0.515	0.460
Qwen2.5-Omni-7B-AWQ		ZS	0.769	0.722	0.789	0.739	0.395	0.300	0.635	0.556
		LoRA	0.788	0.729	0.810	0.758	0.463	0.358	0.629	0.546
GPT-Audio-1.5		ZS	**0.917**	**0.918**	**0.928**	**0.929**	**0.663**	**0.629**	**0.867**	**0.860**

The effect of adaptation is extractor-dependent. With the same PhoWhisper-medium ASR under quiet speech, LoRA improves Gemma from 0.873 (ZS) to 0.896 (LoRA) in macro Presence F1, whereas MedGemma decreases from 0.874 to 0.830. Overall, GPT-Audio-1.5 provides strong and

stable field-presence performance, while the best cascaded systems remain competitive and can achieve higher value extraction quality in some configurations.

4.2.2 DIAGNOSTIC DECOMPOSITION

Table 4 shows that domain adaptation is critical for clinical speech recognition. LoRA adaptation reduces the WER of PhoWhisper-medium from 27.41% to 5.97% on quiet speech and from 36.96% to 11.29% on noisy speech, substantially outperforming both PhoWhisper-large (28.98%/38.30%) and the off-the-shelf Whisper-large-v3 (68.54%/67.95%). This suggests that domain adaptation, rather than model scale alone, is essential for recovering clinical speech. Table 5 further quantifies how much these recognition errors affect downstream structuring. Replacing PhoWhisper+LoRA transcripts with reference transcripts improves all twelve model setting combinations, increasing macro Presence F1 by 0.033-0.119 and VQ-ROUGE-L by 0.087-0.259. However, even with recognition errors removed, no configuration exceeds 0.938 macro Presence F1 or 0.827 VQ-ROUGE-L. Thus, ASR errors account for a substantial part of end-to-end degradation, but structured extraction remains an independent source of error.

Table 4: ASR performance measured by word error rate (WER) across four speech recognizers under quiet and noisy clinical speech conditions on the held-out test set.

ASR Model		PhoWhisper-medium	PhoWhisper-large	PhoWhisper-medium + LoRA	Whisper-large-v3 (no Vietnamese fine-tune)
WER (%)	**Quiet**	27.41	28.98	**5.97**	68.54
	Noisy	36.96	38.30	**11.29**	67.95

Table 5: Structured field extraction from reference transcripts using four LLM backbones under different adaptation settings without ASR errors on the clinical benchmark.

LLM	Qwen2.5-7B			MedGemma-1.5-4B			Qwen3.5-9B			Gemma-4-12B		
Metric	ZS	FS	LoRA	ZS	FS	LoRA	ZS	FS	LoRA	ZS	FS	LoRA
PF1-ma	0.71	0.82	**0.86**	0.72	0.92	**0.93**	0.88	0.93	**0.94**	0.92	**0.93**	**0.93**
PF1-mi	0.76	0.83	**0.88**	0.75	0.93	**0.95**	0.93	**0.95**	**0.95**	0.94	0.94	**0.95**
VQ-R	0.43	0.59	**0.69**	0.39	0.76	**0.80**	0.65	0.77	**0.82**	0.76	**0.83**	0.77
VQ-B	0.52	0.70	**0.75**	0.53	**0.91**	0.87	0.79	**0.89**	**0.89**	0.87	**0.92**	0.86

4.2.3 FIELD-LEVEL HETEROGENEITY IN RESIDUAL STRUCTURED EXTRACTION

Results on reference transcripts show that structured extraction errors persist even after removing ASR errors. The best macro Presence F1 reaches only 0.938, indicating that accurate transcription alone is insufficient for reliable mapping into the target schema. Moreover, micro Presence F1 consistently exceeds macro Presence F1 across all 12 configurations, with gaps of 0.007–0.05, suggesting that extraction reliability is uneven across fields.

This heterogeneity is also reflected in the data characteristics. As shown in Figure 3, fields with similar prevalence can contribute substantially different amounts of target tokens. For example, *Gender*, *Procedures*, and *Clinical entities* are present in more than 96% of cases, yet account for only 0.7%, 10.9%, and 16.0% of target tokens, respectively. Thus, field prevalence does not directly reflect a field's contribution to token-level optimization.

Overall, fields differ substantially in their token contribution, whereas macro Presence F1 assigns equal importance to each field. This mismatch motivates the field-normalized training signal introduced in Section 5.2, where we investigate whether Macro Field Loss improves structured-extraction reliability across model backbones.

5 FIELD-BALANCED STRUCTURED FINE-TUNING

5.1 FROM TOKEN-LEVEL IMBALANCE TO FIELD-LEVEL LEARNING

When the relevant information is correctly represented in the transcript, remaining errors may arise from the IE component failing to map that information to the appropriate target fields. For structured

generation with an LLM, let the token-level cross-entropy at position t

$$\mathcal{L}_{\text{tok}} = \frac{1}{T}\sum_{t=1}^{T} \ell_t, \qquad \text{where} \qquad \ell_t = -\log p_\theta\left(y_t \mid y_{<t}, x\right).$$

The standard training objective averages this loss over all T output tokens. Because the objective is averaged at the token level, longer fields contribute more tokens and therefore receive greater weight during optimization. For structured outputs, this induces length-dependent weighting across fields, even though each field represents a distinct component of the target schema.

5.2 Macro Field Loss

Assume that the output contains K fields. For field k, let $\mathcal{I}_k$ denote the set of token positions belonging to that field and $T_k = |\mathcal{I}_k|$. The field-normalized loss is defined as

$$\mathcal{L}_{\text{field}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k, \qquad \text{where} \qquad \mathcal{L}_k = \frac{1}{T_k}\sum_{t\in\mathcal{I}_k} \ell_t.$$

This normalization gives each field equal weight in the field-level objective, regardless of its token length. The final training objective $\mathcal{L}$ combines the standard token-level loss $\mathcal{L}_{\text{tok}}$ with the Macro Field Loss $\mathcal{L}_{\text{field}}$, with α controlling the contribution of the field-level objective: $\mathcal{L} = \mathcal{L}_{\text{tok}} + \alpha\mathcal{L}_{\text{field}}$.

5.3 Results and Ablation Analysis

Motivated by the field-level heterogeneity observed in Section 4.2.3, we investigate whether explicitly balancing field contributions can better align optimization with structured prediction. Table 6 shows that a moderate field-level weight is beneficial, with $\alpha = 0.25$ achieving the highest macro Presence F1 of 0.943 while maintaining value quality close to the best-performing setting.

Table 7 further evaluates Macro Field Loss across four LLM backbones using this setting. Macro Presence F1 improves on three of the four models, with the largest gain on Qwen2.5-7B, from 0.860 to 0.941. Qwen3.5-9B and Gemma-4-12B also improve from 0.938 to 0.943 and from 0.908 to 0.922, respectively, while MedGemma shows a mixed response. Overall, these results suggest that complementing token-level optimization with a field-normalized objective can improve field-level reliability across different model backbones, particularly when schema fields contribute substantially different numbers of target tokens.

Table 6: Ablation of field-level loss weight α for Qwen3.5-9B LoRA fine-tuning.

Method	PF1-ma	PF1-mi	VQ-R	VQ-B
LoRA	0.938	0.952	0.819	0.889
$\alpha = 1.0$	0.937	0.952	0.821	0.894
$\alpha = 0.5$	<u>0.942</u>	<u>0.954</u>	<u>0.830</u>	**0.906**
$\alpha = 0.25$	**0.943**	<u>0.954</u>	**0.830**	<u>0.905</u>
$\alpha = 0.1$	0.941	**0.955**	0.825	0.896

Table 7: Comparison of LoRA and Macro Field Loss across four LLM backbones with $\alpha = 0.25$.

Model	M.	PF1-ma	PF1-mi	VQ-R	VQ-B
Qwen2.5-7B	LoRA	0.860	0.877	0.691	0.754
	Macro	**0.941**↑	**0.955**↑	**0.811**↑	**0.896**↑
MedGemma-4B	LoRA	**0.931**	0.945	**0.797**	0.873
	Macro	0.929↓	**0.946**↑	0.789↓	**0.877**↑
Qwen3.5-9B	LoRA	0.938	0.952	0.819	0.889
	Macro	**0.943**↑	**0.954**↑	**0.830**↑	**0.905**↑
Gemma-4-12B	LoRA	0.908	0.932	0.749	0.818
	Macro	**0.922**↑	**0.933**↑	**0.758**↑	**0.875**↑

6 Conclusion

We introduced ViMed-Struct, a record-grounded dataset and benchmark for Vietnamese clinical speech-to-structured-EMR generation. By evaluating field selection and value quality separately, our results reveal that structured extraction errors persist even when reference transcripts are used. This finding indicates that, although domain adaptation is essential for improving ASR, better speech recognition alone is insufficient to ensure that clinical information is mapped accurately to the target structure. Motivated by this observation and the imbalance in supervision across fields, we proposed Macro Field Loss, which improves macro Presence F1 for most evaluated backbones and demonstrates the benefit of balancing field-level supervision in structured EMR generation.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval and discovery, research ideation and execution, drafting parts of the manuscript, and improving the clarity and presentation of the writing. All AI-assisted content was reviewed and verified by the authors. References suggested by AI tools were checked against their original sources, and the authors independently validated the scientific claims, experimental results, and conclusions presented in this work. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This study was conducted in accordance with established medical research ethics standards and institutional regulations. The clinical records used in this study were retrospectively collected under approval of the Institutional Review Board (IRB), with a formally granted waiver of informed consent due to minimal risk. All clinical records, derived transcripts, structured annotations, and associated research data were fully de-identified before research use. Personally identifiable information (PII) and other sensitive clinical information were removed in accordance with applicable privacy and data protection requirements. The de-identified dataset will be publicly released for non-commercial research purposes under specified usage terms upon paper acceptance.

REPRODUCIBILITY STATEMENT

The main paper describes the ViMed-Struct construction process, clinical schema, case-level data splits, field-aware evaluation protocol, benchmarked systems and adaptation settings, and the proposed Macro Field Loss. The dataset and associated evaluation resources will be released for non-commercial research under the usage terms described in the Ethics Statement.

REFERENCES

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. An investigation of evaluation methods in automatic medical note generation. In *Findings of the association for computational linguistics: ACL 2023*, pp. 2575–2588, 2023.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. Automatic evaluation of healthcare llms beyond question-answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 108–130, 2025.

Yunmo Chen, William Walden, Tongfei Chen, Aaron Steven White, and Benjamin Van Durme. A unified view of evaluation metrics for structured prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12868–12882, 2023.

Jean-Philippe Corbeil, Asma Ben Abacha, George Michalopoulos, Phillip Swazinna, Miguel Del-Agua, Jérôme Tremblay, Akila Jeeson Daniel, Cari Bader, Kevin Cho, Pooja Krishnan, et al. Empowering healthcare practitioners with language models: Structuring speech transcripts in two real-world clinical applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 859–870, 2025a.

Jean-Philippe Corbeil, Asma Ben Abacha, Jérôme Tremblay, Phillip Swazinna, Akila Jeeson Daniel, Miguel Del-Agua, and Francois Beaulieu. Overview of the mediqa-oe 2025 shared task on medical order extraction from doctor-patient consultations. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pp. 11–16, 2025b.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26813–26829, 2025.

Sy Hwang, Katherine S Pitcher, Sue Hyon Kim, Yoonjae Lee, Hayoung K Donnelly, Harsh Bandhey, Andrew J King, Karen O'Connor, Ryan J Urbanowicz, and Danielle L Mowery. Semantica lab at mediqa-synur 2026: Route, extract, and verify–an llm-gated ensemble for parsing nurse dictations. In *The 8th Workshop on Clinical Natural Language Processing (Clinical NLP)@ LREC 2026*, pp. 27–37, 2026.

Miao Jing, Mengting Jia, Junling Lin, Zhongxia Shen, Huan Gao, Mingkun Xu, and Shangyang Li. Beyond classification accuracy: Neural-medbench and the need for deeper reasoning benchmarks. In *International Conference on Learning Representations*, volume 2026, pp. 87125–87148, 2026.

Annessa Kernberg, Jeffrey A Gold, and Vishnu Mohan. Using chatgpt-4 to create structured medical notes from audio recordings of physician-patient encounters: comparative study. *Journal of medical Internet research*, 26:e54419, 2024.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4958–4972, 2021.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. Phowhisper: Automatic speech recognition for vietnamese. *arXiv preprint arXiv:2406.02555*, 2024.

Khai Le-Duc. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 17365–17370, 2024.

Hyunkyung Lee, Jisoo Jung, Jeonguk Lee, Jaehyo Yoo, Wooseok Han, Minkyu Kim, and Gibaeg Kim. Cap: A source-grounded proposition scaffold for faithful clinical dialogue-to-note generation. In *BioNLP 2026*, pp. 572–594, 2026.

Kefei Liu and Meizhu Liu. Synthetic doctor-patient dialogue generation for robust medical asr: A scalable pipeline for vocabulary expansion and privacy preservation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 525–534, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 72145–72169, 2025.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Yi-Jen Shih, Desh Raj, Chunyang Wu, Wei Zhou, SK Bong, Yashesh Gaur, Jay Mahadeokar, Ozlem Kalinli, and Mike Seltzer. Can speech llms think while listening? In *International Conference on Learning Representations*, volume 2026, pp. 72392–72414, 2026.

Swetha Krishna Sriram and Akshitaa Sahoo. Midas_synur at mediqa-synur 2026: A prompting study for clinical observation extraction from nurse dictation transcriptions. In *The 8th Workshop on Clinical Natural Language Processing (Clinical NLP)@ LREC 2026*, pp. 244–250, 2026.

Meng-Han Su, Wei-Chun Wang, Yi-Min Hsu, Shih-Yen Hou, Su-Jung Chuang, and Shih-Sheng Chang. Voice-based structured nursing documentation using automatic speech recognition and large language models: Development and evaluation study. *JMIR nursing*, 9(1):e88567, 2026.

Hanna Suominen, Liyuan Zhou, Leif Hanlen, and Gabriela Ferraro. Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations. *JMIR medical informatics*, 3(2):e4321, 2015.

Hanna Suominen, Liyuan Zhou, Lorraine Goeuriot, and Liadh Kelly. Task 1 of the clef ehealth evaluation lab 2016: Handover information extraction. 2016.

Mohammad Tanhaei. A hybrid large language model framework for structured data entry from code-switched persian clinical speech. *Scientific Reports*, 2026.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

A Canonical Schema and Field Mapping

Table 8 summarizes the major information groups represented in the canonical clinical schema.

Table 8: Major information groups in the canonical clinical schema.

Information Group	**Content**
Demographics	Gender and age.
Admission information	Reason for admission and related admission information.
Disease course	History of present illness and symptom progression.
Medical history	Personal and family medical history when available.
Physical examination	Clinical findings observed during examination.
Clinical investigations	Laboratory results, imaging findings, and other paraclinical information.
Diagnosis	Primary and accompanying diagnoses.
Treatment	Medications, procedures, interventions, and other treatment information.
Clinical course	Clinical progression, complications, and changes during treatment.
Outcome	Treatment outcome or patient status when available.

Because equivalent clinical concepts may appear under different field names across source records, source fields are mapped to their corresponding canonical fields. Representative examples are shown in Table 9.

Table 9: Examples of source-to-canonical field mapping.

Source A	**Source B**	**Canonical Field**
`QuaTrinhBenhLy`	`QUATRINHBENHLY`	`qua_trinh_benh_ly`
`TienSuBanThan`	`TIENSUBANTHAN`	`history.tien_su_ban_than`
`TienSuGiaDinh`	`TIENSUGIADINH`	`history.tien_su_gia_dinh`
`ChanDoan`	`CHUANDOAN`	`diagnosis.chan_doan`
`PhuongPhapDieuTri`	`PHUONGPHAPDIEUTRI`	`treatment.phuong_phap_dieu_tri`

B Clinical Script Generation and Editing Rules

The operational rules used for clinical script generation and editing are summarized in Table 10.

Table 10: Rules used for clinical script generation and editing.

Aspect	Rule
Information inclusion	Only information supported by the corresponding canonical record is included; fields with null values are omitted.
Clinical ordering	Information is organized according to the progression of a clinical case rather than the storage order of fields in the schema.
Verbalization	Structured values are converted into natural clinical expressions while preserving their original meaning.
Length	Narratives are constrained to approximately 250-2,000 characters.
De-identification	Patient-identifiable information is removed or replaced with generalized descriptions.
Storage terminology	Terms reflecting database or storage conventions rather than clinical communication are removed.
Sentence editing	Short sentences may be merged, and phrases or sentences may be reordered when clinical meaning and temporal relationships are preserved.
Naturalization	Wording and transitions may be adjusted to better reflect natural spoken clinical communication.
Content preservation	Editing must not introduce new clinical events, diagnoses, treatments, outcomes, or other information unsupported by the corresponding record.

A typical narrative follows the clinical progression from admission and presenting symptoms to examination, investigations, diagnosis, treatment, and subsequent clinical course when available. This ordering serves as a general guideline rather than a rigid template, since the available information varies across cases.