Towards Error-Free EHRs: Reasoning-Intensive Consistency Verification Between Clinical Notes and Structured Tables in Electronic Health Records

Anonymous Authors[1]

Abstract

Data consistency between unstructured clinical notes and structured tables in Electronic Health Records (EHRs) is essential for patient safety. However, existing work on note–table consistency verification mainly relies on surface-level matching of numeric values or simple events. Such approaches fail to capture the reasoning underlying real-world EHR documentation, including clinical interpretation, event relations, and temporal changes. To address this gap, we introduce `EHR-ReasonCon`, a reasoning-intensive benchmark for note–table consistency verification. Built on MIMIC-III with expert-guided annotations, it comprises 8,048 entities and provides high-quality ground-truth labels. Our evaluation using expert-validated LLM-as-a-judge metrics reveals the challenging nature of this task; even CheckEHR, the current state-of-the-art in consistency checking, struggles to perform effectively on this benchmark.

1. Introduction

Within Electronic Health Record (EHR) systems, patient information is documented through two primary modalities: structured tables (e.g., vital signs, prescriptions) and unstructured clinical notes (e.g., physician notes) (Seinen et al., 2024). Clinicians combine objective findings from structured data with contextual information from notes to guide diagnostic and therapeutic decisions, which are then recorded back into the EHR. Due to this interdependence, reliability and consistency between these data types are critical. However, in practice, discrepancies frequently arise from administratively driven system architectures and documentation practices (Payne et al., 2018; Villa & Cabezas, 2014), potentially compromising patient safety and creating legal risks (Demsash et al., 2023; Tsou et al., 2017).

Detecting these discrepancies is therefore essential, yet manual verification is impractical due to its time and cost demands. This has motivated the development of automated approaches. Prior work has investigated inconsistencies between clinical notes and structured tables, primarily focusing on specific domains such as allergies or medications (Li et al., 2015; Lo et al., 2022; Rinott et al., 2012). More recently, EHRCon (Kwon et al., 2024) extended this line of work to relational databases. However, these approaches rely on *surface-level* verification, such as checking whether numerical values (e.g., WBC 10.0) or discrete events (e.g., administration of vancomycin) mentioned in clinical notes are also recorded in structured tables. While such approaches provide a useful starting point, they often fail to capture the contextual and nuanced nature of real-world clinical documentation.

Real-world EHR documentation fundamentally requires advanced reasoning for accurate note–table consistency verification beyond surface-level alignment. For example, clinical notes often describe interpreted patient states, whereas structured tables record the underlying measurements (see Figure 1-(1)) (Gao et al., 2024; Raghavan et al., 2014). Verifying these statements therefore requires assessing whether the measurements satisfy the clinical criteria supporting the interpretation. Moreover, clinical notes often describe relationships among multiple clinical events (see Figure 1-(2)). Verifying such statements requires checking whether these event relationships are consistently supported by structured records, rather than validating individual table entries in isolation (Khetan et al., 2022; Wang et al., 2018). Furthermore, clinical notes often describe changes in patient status over time and subsequent interventions (see Figure 1-(3)). Verifying such statements requires assessing trends, time spans, and corresponding treatments rather than relying on a single time point in structured tables (Pan et al., 2020; Yu et al., 2024). However, existing approaches fall short in capturing these reasoning-intensive aspects of note–table consistency.

To address these challenges, we introduce `EHR-ReasonCon`, a *reasoning-intensive* consistency verification benchmark built on MIMIC-III (Johnson et al.,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

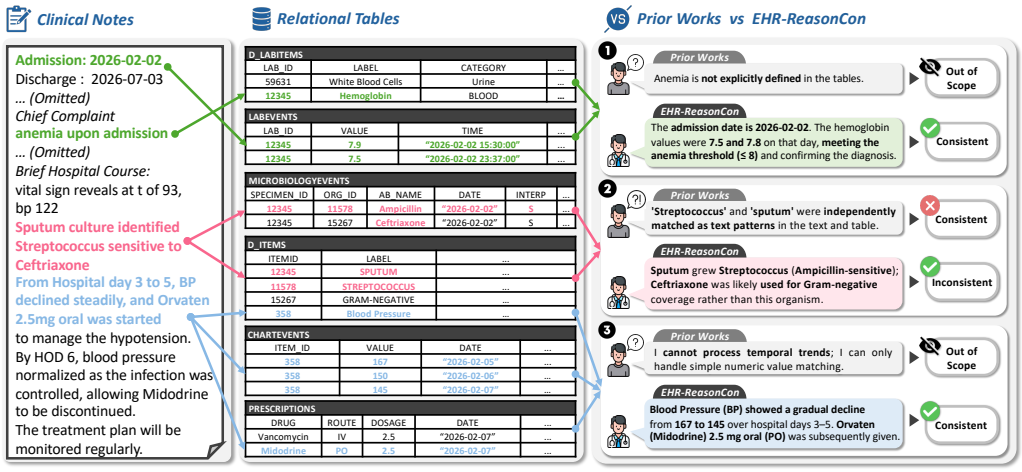

Figure 1. Overview of reasoning-intensive note–table consistency verification. The examples highlight the need for reasoning-intensive verification that goes beyond surface-level alignment between clinical notes and structured tables.

2016)[1]. It evaluates the consistency of 8,048 clinical entities extracted from 105 clinical notes against structured tables. The dataset was constructed using an annotation protocol developed with clinical practitioners, with labeling performed by annotators familiar with EHRs. To support *reasoning-intensive* verification, annotators used eight table-exploration tools and consulted clinical knowledge sources when needed, with final adjudication by physicians to ensure reliable ground truth.

Experimental results further demonstrate the challenging nature of `EHR-ReasonCon`. Notably, even CheckEHR (Kwon et al., 2024), the current state-of-the-art framework for consistency checking, struggles to perform effectively on our benchmark. This limited performance underscores the difficulty of achieving human-level reasoning over structured clinical data and highlights a significant gap in current automated verification approaches.

2. Related Works

Discrepancies between clinical notes and tables have long been recognized as a critical issue that can lead to medical errors (Kwon et al., 2024; Li et al., 2015; Lo et al., 2022; Rinott et al., 2012). Early studies on consistency checking primarily focused on reconciliation within specific domains to align information across disparate data sources. For example, (Rinott et al., 2012) detected inconsistencies in sarcoma discharge summaries using an ensemble of classifiers, (Li et al., 2015) proposed a hybrid ML and rule-based approach for pediatric medication discrepancies, and (Lo et al., 2022) applied NLP methods to reconcile allergy information between clinical notes and structured lists. However, these approaches typically relied on extracting coded entities from notes and comparing them with structured tables, without defining consistency verification as a general task or releasing datasets. To address this limitation, EHRCon (Kwon et al., 2024) introduced a benchmark for verifying consistency between clinical notes and relational databases, constructed on MIMIC-III (Johnson et al., 2016). The dataset includes manual annotations linking entities in clinical notes to entries in multiple tables via SQL query execution. However, EHRCon performs verification in a *surface-level* manner, assessing whether specific values or simple events in notes match structured records. In contrast, `EHR-ReasonCon` introduces a *reasoning-intensive* benchmark for assessing note–table consistency.

[1]MIMIC-III is a publicly available, de-identified EHR database containing clinical data associated with over 40,000 ICU patients at Beth Israel Deaconess Medical Center.

3. EHR-ReasonCon

`EHR-ReasonCon` is a high-quality benchmark for context-aware consistency verification, built from 105 clinical notes across three note types: discharge summaries, physician notes, and nursing notes. The dataset contains 8,048 annotated entities linked to 14 tables in MIMIC-III. Table 1 summarizes statistics of `EHR-ReasonCon`. Figure 2 shows the annotation process, and the steps involved are described below.

Stage 0: Pre-Annotation Setup The goal of this stage is to establish the annotation protocol and tools needed to construct a high-quality benchmark reflecting clinical contexts. The protocol, developed with medical practitioners, specifies how narrative expressions in clinical notes are mapped to structured EHR fields and provides criteria for interpreting temporal trends, handling ambiguous clinical judgments, and ensuring annotation consistency (see Appendix A for detailed protocol). However, real-world EHR data contain edge cases not fully anticipated by predefined

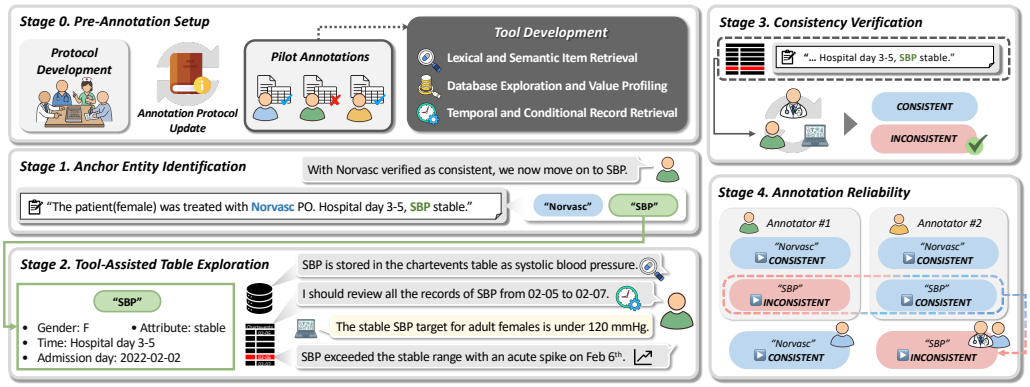

Figure 2. Overview of the reasoning-intensive annotation process for note–table consistency verification. The pipeline includes protocol and tool development (Stage 0), anchor entity identification (Stage 1), tool-assisted evidence retrieval (Stage 2), and entity-level consistency verification (Stage 3). Finally, a dataset-level quality control step (Stage 4) is performed to resolve disagreements and refine annotations after all entities are annotated.

Note Type	Entity		Labels		Note	
	Mean	Total	Con	Incon	Total	Mean Length
Discharge	92.08	3,499	2,042	1,457	38	2,789
Physician	86.15	2,843	2,388	455	33	1,859
Nursing	50.18	1,706	1,418	288	34	1,111
Total	76.65	8,048	5,848	2,200	105	1,953

Table 1. Dataset statistics of EHR-ReasonCon.

guidelines. To address this, we conducted a pilot annotation study to refine the protocol. During the pilot phase, we analyzed how annotators searched for relevant evidence in structured EHR data and formalized recurring search patterns into modular functionalities. This process produced eight table-exploration tools (see Appendix B for further details) that support efficient exploration of complex EHR databases.

Stage 1: Anchor Entity Identification The goal of this stage is to identify anchor entities in clinical notes that correspond to items in structured tables. These anchors serve as entry points for exploring records in tables such as medications and diagnoses.

Stage 2: Tool-Mediated Table Exploration The goal of this stage is to retrieve structured evidence from EHR tables corresponding to the anchor entities. Annotators review the anchor entities, their attributes (e.g., value, procedure location), and their temporal context to understand the clinical trends described in the notes. Based on this understanding, they use the predefined tools (in **Stage 0**) to check the structured EHR tables and identify records that support the narrative content of the clinical notes. The retrieved records are then used as structured evidence for assessing consistency.

Stage 3: Consistency Verification The goal of this stage is to verify whether the narrative content associated with each anchor entity is supported by the structured evidence. Anchor entities are labeled as CONSISTENT if the corresponding information in the clinical notes is supported by structured records; otherwise, they are labeled as INCONSISTENT. To ensure reliable labeling, annotators perform context-aware analysis involving temporal reasoning, commonsense reasoning, and medical interpretation. When necessary, they consult established medical references[2] or physicians.

Stage 4: Annotation Reliability The goal of this stage is to ensure the integrity and reliability of the dataset through a multi-step verification process. The annotation was conducted by eight trained annotators. For each clinical note, two annotators were randomly assigned to independently perform the annotation and then resolve disagreements through mutual reconciliation. For complex clinical judgments, annotators consulted medical professionals. After the initial annotations were completed, an independent reviewer conducted a review of all 105 clinical notes to ensure dataset-wide consistency. The inter-annotator agreement for NER and consistency labeling was 0.897 and 0.888, respectively.

4. Experiments

4.1. Experimental Setting

We split the 105 clinical notes into 83 for the test set and 22 for the validation set. The main experiments were conducted on the test set, and the validation set was used to develop the baseline.

Baseline CheckEHR (Kwon et al., 2024) is the state-of-the-art framework for verifying consistency between clinical notes and structured tables. CheckEHR is an LLM-based eight-stage pipeline that extracts entities from clinical notes

[2]UpToDate, MedlinePlus, Cleveland Clinic, Mayo Clinic

Method	Base LLM	Lenient						Harsh					
		Discharge		Physician		Nursing		Discharge		Physician		Nursing	
		Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec
CheckEHR	MedGemma 27B	17.95	55.19	6.60	35.09	5.56	32.54	16.41	43.23	6.10	32.71	4.80	26.09
	GPT-OSS 20B	29.40	60.19	24.91	39.23	24.00	55.85	26.50	50.70	23.00	35.30	22.20	47.20
	Qwen3 32B	26.05	62.57	27.55	**48.64**	24.05	58.10	23.04	54.02	25.31	**43.95**	22.12	52.14
	Gemini 2.5 Flash	**48.80**	**65.14**	**28.05**	35.86	**38.07**	**64.48**	**44.89**	**55.01**	**26.38**	30.61	**35.16**	**55.59**

Table 2. Performance comparison across different base LLMs. Values are reported as Recall (Rec) and Precision (Prec) under Lenient and Harsh evaluation.

and generates SQL queries to validate them against tables. To the best of our knowledge, it is the only framework that performs entity extraction and table-based consistency checking, and thus serves as our primary baseline.

Base LLMs We use four LLMs as base models for CheckEHR: Gemini 2.5 Flash (Comanici et al., 2025), a proprietary model known for strong reasoning performance with high cost efficiency; Qwen3-32B (Yang et al., 2025) and GPT-OSS 20B (Agarwal et al., 2025), open-source reasoning models; and MedGemma 27B (Sellergren et al., 2025), an open-source model specialized for medical-domain tasks.

4.2. Evaluation

Metrics We evaluate the frameworks at the entity level. For each note, we compute Recall, Precision, and F1 and report the average scores. Recall is defined as the number of correctly classified entities divided by the number of ground-truth entities in the note. Precision is defined as the number of correctly classified entities divided by the number of entities recognized by the framework.

LLM-as-a-judge Entity-level evaluation is typically conducted via direct matching between framework outputs and gold annotations. However, this approach is unreliable for our task due to two main reasons. First, discrepancies in entity span boundaries and granularity can lead to mismatches even when the underlying clinical meaning is equivalent. For example, in the phrase *"lung sounds: clear, no crackles,"* a human annotator may treat *"clear"* and *"no crackles"* as separate verifications for the entity *"lung sounds,"* whereas a framework may extract *"lung"* as the entity and associate both values jointly or separately. Such discrepancies in span boundaries and granularity make exact-match evaluation challenging. Second, outputs that deviate from gold annotations can still be clinically valid. For instance, *"No rash"* may be mapped either to the table item *"Skin integrity"* with the value *"intact"* or to *"Rash"* with the value *"None."* Although these representations differ structurally, both are clinically sound, indicating that strict agreement alone is insufficient for evaluation. To address these issues, we adopt an LLM-as-a-judge evaluator based on Gemini 2.5 Pro (Comanici et al., 2025), which assesses clinical alignment beyond exact matching. We define two complementary evaluation settings to capture both exactness and clinical validity: *Harsh* and *Lenient*. The *Harsh* setting evaluates exact agreement between classification results and gold annotations, while the *Lenient* setting accepts clinically plausible variations even when they diverge from gold annotations. To validate the reliability of the LLM-based judgments, a subset of evaluation samples was independently reviewed by human annotators. The *Harsh* setting achieved 99.46% agreement with author annotations, whereas the *Lenient* setting, involving subjective clinical judgment, reached 95.35% agreement among four practitioner.

4.3. Results

Even when instantiated with different LLMs, CheckEHR achieves F1 scores only in the 50s, underscoring the inherent difficulty of the task. This suggests that current LLM-based approaches remain limited in performing human-level reasoning over structured clinical tables and require more advanced frameworks for effective table reasoning.

Performance is largely driven by the reasoning capability of the underlying LLM. Notably, despite being specialized for medical-domain tasks, MedGemma 27B consistently underperforms all general-purpose reasoning models across our experiments. This indicates that the task requires not only domain-specific medical knowledge but also strong structured reasoning over tabular data, where general reasoning ability is a more critical factor.

5. Conclusion

In this work, we introduce `EHR-ReasonCon`, a reasoning-intensive benchmark for verifying consistency between clinical notes and structured tables in EHRs. Our analysis shows that existing LLM-based approaches struggle to perform effectively on this task, highlighting the gap between current capabilities and the level of reasoning required in real-world clinical settings. We hope this benchmark will serve as a foundation for future research on reasoning over structured clinical data and inspire the development of more robust and clinically reliable verification systems.

Impact Statement

This paper introduces a benchmark to improve the reliability of Electronic Health Records (EHRs). While our work aims to advance machine learning in healthcare, we recognize the sensitivity of clinical data. To mitigate ethical risks, we utilized de-identified, public datasets (MIMIC-III) and strictly followed established clinical annotation protocols. We believe our work contributes to safer AI-assisted clinical documentation without introducing immediate ethical concerns.

References

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Demsash, A. W., Kassie, S. Y., et al. Health professionals' routine practice documentation and its associated factors in a resource-limited setting: a cross-sectional study. *BMJ Health & Care Informatics*, 30(1):e100699, 2023. doi: 10.1136/bmjhci-2022-100699.

Gao, Y., Mahajan, D., Uzuner, Ö., and Yetisgen, M. Clinical natural language processing for secondary uses. *Journal of Biomedical Informatics*, 150:104596, Feb 2024. doi: 10.1016/j.jbi.2024.104596.

Johnson, A., Pollard, T., and Mark, R. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Khetan, V., Rizvi, M. I., Huber, J., Bartusiak, P., Sacaleanu, B., and Fano, A. Mimicause: Representation and automatic extraction of causal relation types from clinical notes. In *Findings of the association for computational linguistics: ACL 2022*, pp. 764–773, 2022.

Kwon, Y., Kim, J., Lee, G., Bae, S., Kyung, D., Cha, W., Pollard, T., Johnson, A., and Choi, E. Ehrcon: Dataset for checking consistency between unstructured notes and structured tables in electronic health records. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Li, Q., Spooner, S. A., Kaiser, M., Lingren, N., Robbins, J., Lingren, T., Tang, H., Solti, I., and Ni, Y. An end-to-end hybrid algorithm for automated medication discrepancy detection. *BMC Medical Informatics and Decision Making*, 15(1):37, 2015. doi: 10.1186/s12911-015-0160-8. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC4427951/.

Lo, Y.-C., Varghese, S., Blackley, S., Seger, D. L., Blumenthal, K. G., Goss, F. R., and Zhou, L. Reconciling allergy information in the electronic health record after a drug challenge using natural language processing. *Frontiers in Allergy*, 3:904923, 2022. doi: 10.3389/falgy.2022.904923.

Pan, X., Chen, B., Weng, H., Gong, Y., and Qu, Y. Temporal expression classification and normalization from chinese narrative clinical texts: Pattern learning approach. *JMIR Med Inform*, 8(7):e17652, Jul 2020. ISSN 2291-9694. doi: 10.2196/17652. URL https://medinform.jmir.org/2020/7/e17652.

Payne, T. H., Alonso, W. D., Markiel, J. A., Lybarger, K., Lordon, R., Yetisgen, M., Zech, J. M., and White, A. A. Using voice to create inpatient progress notes: effects on note timeliness, quality, and physician satisfaction. *JAMIA open*, 1(2):218–226, 2018.

Raghavan, P., Chen, J. L., Fosler-Lussier, E., and Lai, A. M. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? In *AMIA Joint Summits on Translational Science Proceedings*, volume 2014, pp. 218–223, Apr 2014.

Rajkomar, A., Loreaux, E., Liu, Y., Kemp, J., Li, B., Chen, M.-J., Zhang, Y., Mohiuddin, A., and Gottweis, J. Deciphering clinical abbreviations with a privacy protecting machine learning system. *Nature communications*, 13(1):7456, 2022.

Rinott, R., Torresani, M., Bertulli, R., Goldsteen, A., Casali, P., Carmeli, B., and Slonim, N. Automatic detection of inconsistencies between free text and coded data in sarcoma discharge letters. In *Studies in Health Technology and Informatics*, volume 180, pp. 661–666, 2012. doi: 10.3233/978-1-61499-101-4-661.

Seinen, T. M., Kors, J. A., van Mulligen, E. M., and Rijnbeek, P. R. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27, 2024. URL https://api.semanticscholar.org/CorpusID:274208819.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Tsou, A. Y., Lehmann, C. U., Michel, J., Solomon, R., Possanza, L., and Gandhi, T. Safe practices for copy and paste in the ehr: Systematic review, recommendations, and novel model for health it collaboration. *Applied Clinical Informatics*, 8(1):12–34, 2017. doi: 10.4338/ACI-2016-09-R-0150.

Villa, L. B. and Cabezas, I. A review on usability features for designing electronic health records. In *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 49–54, 2014. doi: 10.1109/HealthCom.2014.7001812.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu, H. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.11.011. URL https://www.sciencedirect.com/science/article/pii/S1532046417302563.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yu, D., Stidham, R. W., and Vydiswaran, V. G. V. A systematic temporal extraction pipeline for medical concepts in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2023, pp. 1314–1323, Jan 2024.

A. Protocol

A.1. Task

Identify entities within clinical notes and compare their associated values with actual tables to detect discrepancies between the notes and the tables. For efficient labeling, we used Streamlit[3], and a screenshot of the interface is shown in Figure 3.

A.2. Annotation Protocols

A.2.1. Definition of Clinical Note

To ensure accurate labeling, it is important to understand the types of clinical notes used in our study. These include the Discharge Summary, Physician Note, and Nursing Note.

Discharge Summary The Discharge Summary is written at the time of the patient's discharge, summarizing the events during their hospitalization. It may include information from before admission, such as past medical history, as well as details from after discharge. For example, information about "admission medications" may be recorded to continue treatment for medications the patient was taking prior to hospitalization.

Physician Note The Physician Note is written by the physician during daily rounds. It describes the patient's condition and outlines the next steps for diagnosis and treatment.

Nursing Note The Nursing Note is written by a nurse and documents the patient's condition. These notes are often recorded multiple times a day to provide ongoing updates.

A.3. Definition of Entity

The goal of entity extraction in this study is to extract all entities that can be matched to the item names from 14 clinical tables, including D_ITEMS, DIAGNOSES_ICD, D_ICD_DIAGNOSES, D_ICD_PROCEDURES, D_LABITEMS, INPUTEVENTS_MV, CHARTEVENTS, MICROBIOLOGYEVENTS, INPUTEVENTS_CV, OUTPUTEVENTS, LABEVENTS, PROCEDUREEVENTS_ICD, PRESCRIPTIONS, and PROCEDUREEVENTS_MV.

Entity Group Considerations In particular, when general drug categories such as "Antibiotics" or "Beta-blockers" or test panels like "ABG" or "Chem-7" appear in the clinical notes, they are expanded into detailed items for analysis. For example, if "Chem-7" is mentioned, the entity values from the LABEVENTS table, such as Sodium, Potassium, Chloride, Bicarbonate, Blood Urea Nitrogen, Creatinine, and Glucose, should be compared with each respective value. To find these detailed items, searches are limited to four established sources: MedlinePlus [4], Cleveland Clinic [5], Mayo Clinic [6], and UpToDate [7]. If the search results are insufficient or additional medical knowledge is required, please consult with a physician.

Entity Mapping in Clinical Notes In clinical notes, information is often written as free text, which means the entities might not always be clearly listed or categorized in the tables. For example, a note that mentions "headaches" might correspond to different entries in the database, such as "pain location" or "type of pain." Even if these entries aren't explicitly labeled, it's essential to match the data to the correct entity in the tables whenever possible. To map entities accurately, it's important to understand how each entity is stored in the database. Since information in clinical notes can sometimes be incomplete or unclear, you need to interpret it carefully and know how to connect it correctly. If something is unclear, it's important to consult with a physician to ensure the mapping is done correctly.

Entity Scope To prevent incorrect discrepancies, we do not extract the following types of information as entities:

- **Past Information**: The past medical history, family history, previous hospital admissions, emergency room visits, and

[3] https://streamlit.io/
[4] https://medlineplus.gov/
[5] https://my.clevelandclinic.org/
[6] https://www.mayoclinic.org/
[7] https://www.uptodate.com/contents/search

medications administered during hospitalization recorded in clinical notes are all considered past information. Since this information is not stored in the tables, it is not extracted as entities.

- **Future Plans**: Information such as 'discharge plans' or 'next steps to be taken' mentioned in discharge summaries or physician notes are categorized as future plans. These plans may not be executed due to changes in the patient's condition, and therefore, are not considered discrepancies if not carried out. Future plans are not extracted as entities, including those recorded as 'plan' in nursing notes.

- **Information Unrelated to Database**: If the information recorded in clinical notes is not linked to a specific record or item in the tables, it is not extracted as an entity. Additionally, information such as transfer details, which are not related to the tables we handle, or items that only mention total amounts without specifying what those amounts refer to, are also not extracted.

Insurance-related Entities Items recorded under "Discharge Diagnoses" or "Major Surgical/Invasive Procedures" in the discharge summary are typically written for insurance claim purposes by hospitals. Entities listed under these items should be limited to discrepancy checks related to insurance claims, specifically for ROCEDUREEVENTS_ICD, DIAGNOSES_ICD, D_ICD_DIAGNOSES, and D_ICD_PROCEDURES.

A.4. Definition of Entity Attribution

Time Structured tables in EHR are essentially time-series data, making the timing of events crucial for understanding a patient's clinical journey. When reviewing clinical notes, it is important to determine the time an event occurred. For example, if a note says "3 days after surgery," the surgery date should be extracted, converted into a standard time format, and then used to calculate the exact date that is 3 days after the surgery. The key is to carefully consider the context of the note and extract the most accurate time possible. While the guidelines below are helpful, the main goal is to interpret the context of the note and determine the most appropriate timestamp.

- **Handling Ambiguous Time Information**: If explicit time information is not provided, the time is inferred based on the note's nature. For example, a discharge summary typically summarizes the patient's admission and discharge records. If the exact time is unclear, the entities in the discharge summary should be checked against the EHR records within the patient's hospitalization period, and if they match, it can be considered consistent. Physician or nursing notes are usually written daily, so in this case, the charting date is treated as the event date. That is, if the entity information is recorded in the EHR on the charting date, it can be considered consistent.

- **Mapping Time Based on Admission and Discharge Dates**: When specified as "admission date" and "discharge date," the time expression may vary depending on the physician. Therefore, a reasonable medical range should be applied, for instance, one day before admission and one day before discharge. Consistency checks should be conducted within this time frame.

- **Medication Timing**: For medication-related tables, such as inputevents or prescription tables, both the start and end times are recorded. If both the start and end times are explicitly provided, these should be checked against the tables' startdate or starttime and enddate or endtime fields to ensure they match accurately. If a duration is recorded in the clinical note, verify that the medication was administered correctly during the specified period.

- **Verifying Time-Related Terms in EHR**: When terms like "morning" or "midnight" are used in a clinical note, verify whether the records exist accurately within the corresponding time frame.

- **Note**: While these guidelines should be followed, the main goal is to understand the context of the note and extract the most accurate timestamp possible.

A.5. Discrepancy Detection Process

The core of this task is to compare the entities extracted from clinical notes with the table records to check for discrepancies. To do this, the following three main approaches can be used:

Matching Based on Common Sense Even if an entity in the clinical note does not exactly match the one in the tables, it can still be considered a match based on common sense. For example, the clinical note may mention "hand," while the tables record "finger." Since both refer to the same part of the body, they can be considered a match.

Matching Based on Medical Knowledge Clinical notes describe a patient's condition in free text, while the tables use a standardized format. For instance, if the note says 'edema 2+,' it might be recorded in the tables as 'palpable edema.' In such cases, it is essential to understand the medical meaning of these terms (e.g., 2+ and palpable) and refer to authoritative sources such as UpToDate, MedlinePlus, Mayo Clinic, and Cleveland Clinic to make an accurate match. If these sources do not provide sufficient information, it is important to consult with a physician for clinical interpretation.

Exact Matching Some clinical note entries may be copied from the tables based on medical practices. In this case, the entities in the clinical note may match the table records exactly. For example, if the note mentions a WBC count of 10.0, the same value may appear in the tables. However, EHR system errors could lead to discrepancies, so it's crucial to carefully check for consistency.

A.6. Single or Multiple Row Matching

Single Row Matching Some entities can be compared with a single row in the tables to check for discrepancies. This applies to events that occur at a specific point in time. For example, if the note says "WBC stable," and the tables show a stable value, it can be considered a match. In this case, as the event happens at a single point in time, finding this record in the tables once would be enough to confirm the match. However, time should also be carefully considered in these cases.

Multiple Row Matching Clinical notes often describe a patient's condition over time. When this happens, the trend across multiple records in the tables needs to be checked. For instance, if the note says, "No fever from hospital day 3 to 5, but fever started on day 4," the table records must show no fever from days 3-5, and the fever should be recorded starting from day 4.

- **Note**: Values like blood pressure (BP) can be represented as follows: 60/100(80)-100/140(120). In this notation, the value in parentheses represents the mean value, while the part before the hyphen indicates the minimum blood pressure, and the part after the hyphen represents the maximum blood pressure range. This notation is used when recording a patient's blood pressure measurement to provide both the average value along with the minimum and maximum values. When comparing with table data, the mean value should be searched as the mean BP in the table, while the minimum and maximum values should be considered as the BP measurement range.

A.7. Example

- WBC 20.0 *
  - At this point, you need to confirm that the numeric value of WBC is 20.0, and that the FLAG column correctly shows the value as "ABNORMAL" in the table.

- Anemia on admission
  - Here, you need to check that the hemoglobin value (valuenum) listed in the table falls within the range indicating anemia, from the day before admission to the day after admission.

- Sputum culture identifying Streptococcus sensitive to Cefazolinex.
  - In this case, the table should show that the specimen is "sputum" and the organism is "Streptococcus". Additionally, you need to confirm that the sensitivity test result (interpretation) for "Cefazolinex" (ab_name) shows "sensitive."

A.8. Consistency Check

Labeling should be strictly limited to consistent and inconsistent. A label of consistent indicates that all information described in the clinical note is accurately aligned with the table. If even a single column contains conflicting information or if there is no supporting record in the table, the case should be labeled as inconsistent. In addition, please specify the rows in the table that serve as evidence for your consistent or inconsistent decision. Notes for Consistency Checking:

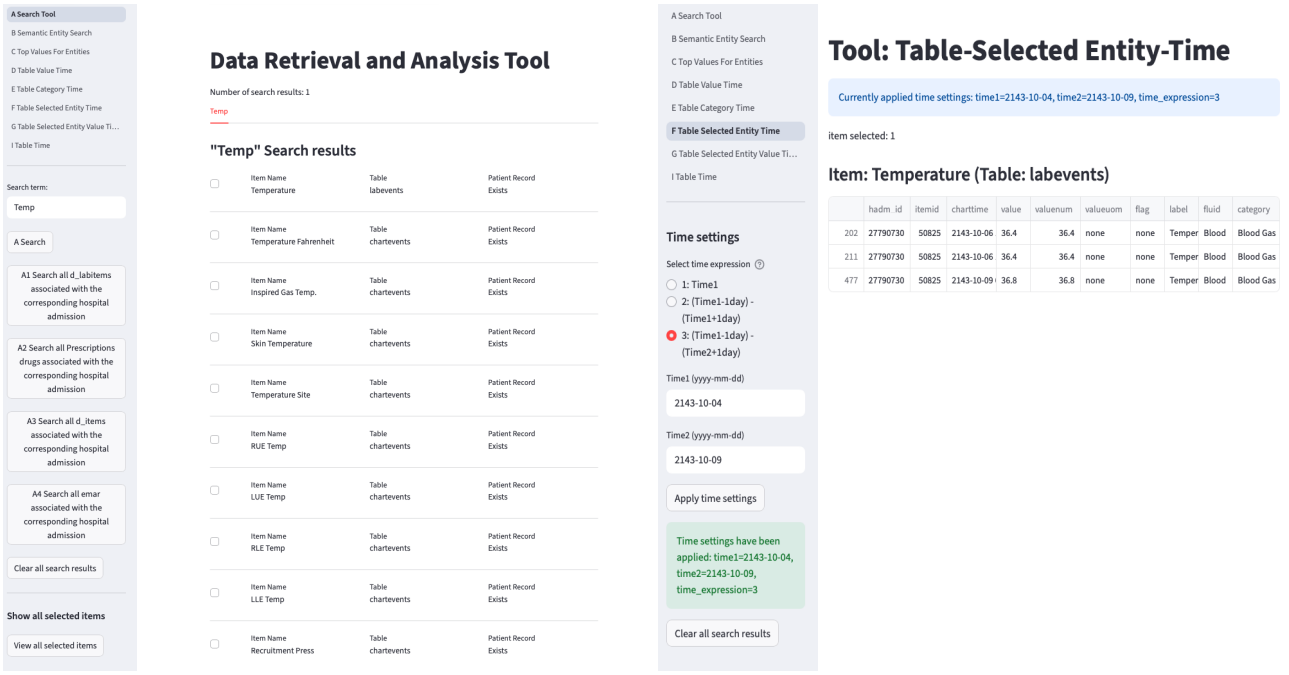

Figure 3. A screenshot of Streamlit provided to labelers for annotation

- **Use of medical knowledge**: If medical knowledge is used to make a judgment, please provide the corresponding reference.

- **Use of commonsense knowledge**: If commonsense knowledge is used, please explicitly state that commonsense knowledge was applied. For example, if the note states that the patient injured their right arm, while the table records an injury to the right finger, this may be considered consistent based on commonsense knowledge.

B. Table-Exploration Tools

This process produced eight table-exploration tools that support efficient exploration of complex EHR databases and are organized into three functional categories:

B.1. Entity-to-Table Item Alignment

The tools in this category support the alignment of entities mentioned in narrative notes with corresponding items in structured EHR tables. The same clinical information may appear under different names or levels of abstraction (e.g., "White Blood Cells" and "WBC"), so these tools retrieve potentially relevant table items based on both lexical similarity and conceptual relatedness.

Lexical_Search This tool uses an N-gram–based search approach to identify entities in a patient dictionary table that are morphologically similar to a given query entity. Rather than relying solely on simple string matching, it incorporates the C4-WSRS medical abbreviation dataset (Rajkomar et al., 2022), allowing abbreviations such as "WBC" or "BP" to retrieve their full forms, "White Blood Cell" and "Blood Pressure." This enables more accurate retrieval by accounting not only for surface-level text similarity but also for abbreviation expansion.

Semantic_Search This tool supports semantic search to capture similarities that are difficult to detect with lexical methods alone. For example, the term "alert," which describes a patient's condition, is conceptually related to "level of consciousness," but this relationship may not be identified through N-gram matching. This tool addresses this limitation by considering context and meaning, linking expressions that differ lexically but represent the same or similar clinical concepts.

B.2. Database Exploration and Value Profiling

The tools in this category support exploration of the EHR database schema and content. Since clinical concepts may be distributed across heterogeneous tables, the tools support exploration of relevant table groups and summarize typical values for each item, enabling annotators to quickly interpret the role of different fields.

Get_Item_Value_Distribution This tool provides insight into the distribution of values associated with specific entities within tables. It allows users to view the top K most frequent values for each entity, helping to characterize the nature of the data. In this study, K was set to 10. For instance, if "SBP" values frequently appear as 110, 120, or 130, this indicates that the entity represents continuous numerical data. This tool enables users to determine whether an entity is numerical, categorical, or follows a specific pattern, and to assess whether the retrieved entity appropriately reflects the information recorded in clinical notes.

Analyze_Category_Trend This tool helps users understand the structure of entities across multiple tables in a database. Since entities are distributed across tables based on their characteristics, it is not always intuitive to determine where a specific entity resides. This tool analyzes how each table is organized into categories and what entities belong to each category, providing insight into the location and context of entities. For example, the chartevents table may include categories such as Labs and General. The Labs category contains items like "anion gap" or "CK-MB," while the General category includes entities related to consciousness, such as "Level of Consciousness" or "Oriented." This helps users more efficiently identify the appropriate table for a given entity.

B.3. Temporal and Conditional Record Retrieval

The tools in this category support verification of clinical statements that involve temporal changes or specific conditions. The tools allow annotators to retrieve records from structured tables based on time windows and value constraints, enabling inspection of whether the structured data support trends or events described in clinical notes.

Get_Item_Status_History This tool enables users to examine entity information over time. Time can be specified in three different ways to refine the search. First, users can query based on an exact standard timestamp, such as "06-24," to check whether records exist at a specific point in time. This approach is suitable when explicit time information is directly recorded in the data. Second, this tool supports searches based on time expressions, such as "admission," which are described narratively in clinical notes. Because these expressions do not correspond to precise timestamps and must be interpreted from context, the search is performed by defining a time window around the inferred point, typically extending from one day before to one day after the reference time, in order to retrieve relevant surrounding records. Third, this tool allows searches based on duration. Users can define a start time (time1) and an end time (time2), and for more robust retrieval, the search range is typically expanded to include the period from one day before the start time to one day after the end time, ensuring that all relevant data within the interval are captured.

Get_Item_Value_History This tool supports more fine-grained queries than Get_Item_Status_History. It allows users to specify not only the entity but also its associated values as search conditions. By using operators such as "more," "less," and "between," users can define value ranges more precisely and retrieve data that meet specific criteria.

Analyze_Value_Trend This tool enables the analysis of value trends over time, independent of specific entities. Rather than focusing on values at a single time point, it examines how values evolve across time, capturing patterns such as increases, decreases, or stability. Users can specify a time range to analyze trends within a particular interval, allowing for a more contextual understanding of value changes. In addition to absolute values, this tool considers dynamic characteristics such as the rate of change and variability. It is designed to move beyond entity-specific status queries and instead support a broader understanding of temporal value patterns across the data.

View_General_Timeline This tool also allows users to retrieve all records associated with a specific time point within a selected table. For example, if "2026-05-03" is set as the reference time for the chartevents table, all patient records documented in the chartevents table on that date can be retrieved.