Identifying Genitourinary Symptoms in Medical Notes with Open-Source Large Language Models

Yunbing Bai, MA, Wanting Cui, MA, Joseph Finkelstein, MD, PhD, FAMIA
Arizona Telemedicine Program, the University of Arizona, Tucson, AZ

Introduction

Early identification of disease relies on accurate recognition of signs and symptoms (S&S). S&S not only guide diagnosis, they also strengthen predictive models for urinary cancers and other conditions [1]. Most S&S live in clinicians' free-text notes—rich but unstructured data that can inform diagnosis, risk prediction, and treatment decisions. While ChatGPT can extract S&S, its use on protected health information raises privacy concerns. We therefore evaluate the feasibility of **locally run Meta-Llama models to extract urinary-system S&S from clinical notes and map to ICD-10-CM codes**, aiming to preserve privacy while maintaining high extraction quality.

Methods

We evaluated local LLM extraction of genitourinary signs & symptoms (S&S) and ICD-10 mapping from real clinical text.

Data source & selection
27 urology notes from the MTSamples corpus, manually screened for explicit genitourinary S&S.

Annotation (gold standard)
Clinical experts labeled genitourinary S&S and mapped each to ICD-10-CM codes in the R30–R39 block.

Model
Meta-Llama-3.3-70B (locally run) was chosen for evaluation based on prior performance signals [2,3].

Prompt design
Goal: (i) extract only the exact S&S mentioned in the note; (ii) map each S&S to the most appropriate ICD-10-CM code.

- Iteration I: include the **ICD-10-CM code of R30-R39**, and **exclude denied** or negative symptoms and signs
- Iteration II: based on iteration I, Instructed the model to **avoid assumptions** and infer nothing beyond the text.
- Iteration III: based on iteration II, enforced a **structured response format** for consistency and easy comparison.

TASK: Extract patient's current presented symptoms and signs involving the genitourinary system from the clinical note.

REQUIREMENT:
- Exclude the denied or negative symptoms and signs. List the symptoms and signs and ICD-10 code.
- Only include the symptoms and signs in the following ICD-10 list:

[list of ICD-10 codes with definition, e.g. R31 Hematuria]

CLINICAL NOTE: *[note here]*

Only return a list of symptoms and signs along with their ICD-10 code in the provided list!

TASK: Extract patient's current presented symptoms and signs involving the genitourinary system from the clinical note.

REQUIREMENT:
- **ONLY INCLUDE explicitly mentioned symptoms and signs.**
- Exclude the denied or negative symptoms and signs. List the symptoms and signs and ICD-10 code.
- Only include the symptoms and signs in the following ICD-10 list:

[list of ICD-10 codes with definition, e.g. R31 Hematuria]

CLINICAL NOTE: *[note here]*

Only include the symptoms and signs that directly mentioned in the notes. Do not make any assumption.

TASK: Extract patient's current presented symptoms and signs involving the genitourinary system from the clinical note.

REQUIREMENT:
- ONLY INCLUDE explicitly mentioned symptoms and signs.
- Exclude the denied or negative symptoms and signs. List the symptoms and signs and ICD-10 code.
- Only include the symptoms and signs in the following ICD-10 list:

[list of ICD-10 codes with definition, e.g. R31 Hematuria]

Return a list of extracted symptoms and signs that are directly mentioned with ICD-10 codes and strictly follow the following format and example. Do not return any additional information.

FORMAT: [ICD-10 code 1 - extracted symptoms and Signs 1, ICD-10 code 1 - extracted symptoms and Signs 2]
EXAMPLE: [R31.2 - microhematuria, R35.1 - nocturia]

CLINICAL NOTE: *[note here]*

Figure 1. Prompt iteration. Bold fonts highlights the undated context.

Results

Among 27 annotated clinical notes, we identified **53** genitourinary symptoms and signs, averaging **1.89 ± 1.22** per note. **Hematuria (R31)** was the most frequent S&S. The **assumption-free prompt without structured output** achieved the best performance (**F1 = 0.92** for extraction, **0.89** for ICD-10 mapping).

Table 1. Evaluation of model outputs for both the Symptoms and Signs extraction task and the ICD-10 code mapping task, using three different prompt settings.

	S&S Extraction			ICD-10 Code Generation		
Prompt Setting	**Precision**	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
ICD-10 Code-Based Symptom Extraction	0.68	0.97	0.8	0.71	0.97	0.82
ICD-10 Symptom Extraction with Assumption-Free Constraints	0.89	0.96	0.92	0.85	0.93	0.89
Standardized ICD-10 Symptom Extraction Format	0.73	0.82	0.77	0.71	0.80	0.75

Discussion and Conclusions

Overall performance
- Extraction > ICD-10 mapping in accuracy.
- Best results came from the **assumption-free prompt**.

Prompting insights
- Adding brief **ICD-10 definitions** reduced mislabels
- Biggest failure mode: **inferring non-textual S&S** → false positives, lower precision.
- Emphasizing "**extract only explicitly stated S&S**" improved precision with minimal recall loss.

Output formatting
- A **structured response format** aids automation but occasionally causes formatting errors due to complexity.
- Trade-off: structure vs. stability.

Limitations
- Coding within R30–R39 still prone to ambiguity.
- Small dataset (27 notes) may limit generalizability.

References

1. Harrison H, Thompson RE, Lin Z, et al. Risk Prediction Models for Kidney Cancer: A Systematic Review. Eur Urol Focus. 2021;7(6):1380-1390.
2. Meta AI. Llama 3.3 70B: a large language model. Available via Ollama. 2023;Version 3.3. Retrieved from https://ollama.com/library/llama3.3.
3. Bai Y, Cui W, Finkelstein J. Utilizing Open-Source Large Language Models to Extract Genitourinary Symptoms from Clinical Notes. Stud Health Technol Inform. 2025;328:16-20. doi:10.3233/SHTI250664