MedPromptExtract (Medical Data Extraction Tool): Anonymization and High-Fidelity Automated Data Extraction Using Natural Language Processing and Prompt Engineering

Article:
Roomani Srivastava, Lipika Bhat, Suraj Prasad, Sarvesh Deshpande, Barnali Das, Kshitij Jadhav.
MedPromptExtract (Medical Data Extraction Tool): Anonymization and High-Fidelity Automated Data Extraction Using Natural Language Processing and Prompt Engineering. J Appl Lab Med 2025; 10(4): 793–805. https://doi.org/10.1093/jalm/jfaf034

Guest: Dr. Kshitij Jadhav from the Koita Centre for Digital Health of IIT Bombay, where he heads the Artificial Intelligence in Digital Health Lab.

Randye Kaye: Hello, and welcome to this edition of *JALM* Talk from *The Journal of Applied Laboratory Medicine*, a publication of the Association for Diagnostics & Laboratory Medicine. I'm your host, Randye Kaye. While electronic medical records have revolutionized healthcare, some information can still be difficult to extract. Discharge summaries, which provide comprehensive information of a patient's history, treatment, and post-discharge care instructions, can be valuable sources of data, but their narrative formats make them difficult to analyze. In low- and middle-income countries, health data may be managed in a hybrid format where only partial records are available through electronic medical records.

The July 2025 issue of *JALM* features an article that introduces MedPromptExtract, an automated method to efficiently extract data from discharge summaries while maintaining patient privacy and compliance with ethical standards in healthcare. The authors used prompt engineering and a large language model to extract custom clinical information from free texted discharge summaries of patients having acute kidney injury. The authors proposed that this innovative tool may aid research, clinical decision making, and operational efficiency in healthcare systems.

Today, we're joined by the article's corresponding author, Dr. Kshitij Jadhav. Dr. Jadhav is an assistant professor of the Koita Centre for Digital Health of IIT Bombay, where he heads the Artificial Intelligence in Digital Health Lab. Trained as a medical doctor and neuroscientist, he specializes in multimodal data analysis and applies advanced artificial intelligence and machine learning techniques to digitize and interpret clinical information.

Welcome, Dr. Jadhav. First, how much useful information is still residing in offline formats in hospitals, and why is it important to digitize this medical data?

Kshitij Jadhav: So, let's say that digitization of hospital records in a country like India is pretty challenging, primarily because there are several players in the healthcare sector. So, it is not as centralized as the NHS in the UK, for example. So, hospitals

in India are still trying to transition into electronic health records and electronic medical systems. As a result of which, a lot of our healthcare data is still clinical notes made by doctors when they see the doctor at bedside or at emergency or outpatient centers and so on.

So, most of these clinical notes then ultimately result into-- trickle down into discharge summaries, which are some of the important hospital records that exist in the country, and also a treasure trove of important information about how the patients' condition and recovery has progressed. So, these insights can help us understand what are the treatment routes that we utilize and what are the decision-making pathways that would help the patient recover. Basically, it separates case scenarios that have worked and that have not worked.

So, coming to the second part of your question, you say that it's basically, it's important to digitize in order to leverage this information, which is currently in an unstructured form, but that has to be converted into a structured database to inform further downstream applications, maybe driven by machine learning algorithms such as using referral services or using clinical decisions.

Randye Kaye: All right, thank you. Can you tell us some more about MedPromptExtract, like what is it and how does it work?

Kshitij Jadhav: So MedPromptExtract actually is a solution that we developed as a precursor to one of the problem treatments that we were working on with Kokilaben Hospital in Mumbai. So MedPromptExtract is a fully automated tool, which is designed for extracting high-fidelity clinical data from discharge summaries, especially when these data points are not digitized. So, it has a three-stage process. So without -- and the process is supposed to function in a scenario so that the privacy is not compromised.

So, first we use something called as an EIGEN framework, which is a semi-supervised learning method, which has been developed at IIT Bombay to remove personal identification information like patient's name, patient's id, and the patient -- what the advantage of such a system, the semi-supervised framework that we use, is that it utilizes very less labeled data. You are able to anonymize thousands of records with just ranking labeled data points from the discharge summaries. This is also a focus in my lab actually, that we want to work and develop solutions for resource-constrained scenarios.

Now, after anonymization, the records of MedPromptExtract actually uses simple NLP [natural language processing] algorithms, such as regular expressions, to pass these

discharge summaries and identify information like diagnosis or medications so that they can be stored in a structured format. Even after this now there's a separate section called clinical notes, which is freeform text that is present. And for this we actually utilize large language models to extract nuanced clinical information. For example, was acute kidney injury diagnosed, what was the hospital duration, was there a contrast involved while collecting this while collecting this in the patient? From a free -- we basically utilized open source models for this, which were crafted by specific clinically designed prompts.

Randye Kaye: Thank you. Now, the paper describes your iterative refinement of the prompts using feedback from clinicians. Can you expand a bit on this, and how pivotal is the role of the clinician using these methods?

Kshitij Jadhav: Yeah, this is a very good question. So, the development of prompts in our project were very much an iterative and a collaborative process with clinicians. So, we started with some basic literature review, but also then took a lot of inputs from nephrologists, focusing on certain key features that are important for acute kidney injury. So subsequently what we did is we tested within small batches of 10 discharge summaries at a time and then revised them based on the detailed feedback that clinicians are giving us. So, the role of the clinicians was actually absolutely central. They not only helped us define which clinical features are extremely important to extract, but also helped us validate the model outputs. It must be also understood that the use case that we have described in the paper MedPromptExtract is to extract data from discharge summaries of acute kidney injury and the method -- however, the method that the MedPromptExtract has developed is pretty conditional and prompts can be modified, so for other use cases as well.

So constant support of clinicians is extremely crucial for developing this initial set of prompts. Now the feedback of clinicians was also very important to resolve ambiguity like distinguishing a general mention of rise in creatinine from a clear diagnosis of acute kidney injury. So, we had this very strong collaborative effort with them that we were able to arrive to a defined, well-defined, domain-sensitive prompt that worked without needing additional fine tuning of some foundational language models. Or manually annotating the large dataset, which ultimately actually save a lot of time, valuable time of our clinicians.

Randye Kaye: All right, thank you. So, in the paper, you also emphasized the importance around anonymization and maintaining patient confidentiality. So, what methods did you implement to ensure this data privacy?

Kshitij Jadhav: Like I mentioned previously, we use a homegrown semi-supervised solution, homegrown in the sense that developed in IIT Bombay, which requires very little annotated data. What I would also like to say is that this anonymization process was done within the hospital ecosystem within the medical cost department, no external data sharing was done, so this ruled out any sort of data breach. Also, the time taken to run the semi-supervised process on a thousand records was negligible. And the compute required for it was also not huge. As a result of you were able to do this within the hospital ecosystem. After data extraction, we were able to port that information, this anonymized information, for further processing in a structured format. So that's how we maintained the data privacy.

Randye Kaye: How do you see tools like MedPromptExtract influencing ethical standards or policies regarding medical data management, especially in resource-constrained settings?

Kshitij Jadhav: So, well, the widespread use and advancement of large language models to some extent are democratizing a lot of such processes. But in the medical domain, ethical concerns and data privacy is still a major concern. So this data privacy is central to the functioning of MedPromptExtract, has a built-in anonymization step for faster de-identification of this data. So, we have eventually installed this tool in the hospital itself, ruling out the need for porting any documents from the hospital to outside settings. So, it really has the potential to address this landscape where it comes to ethical medical data management. Speaking of resource-constrained scenarios, one of the biggest advantage is that it supports cost effective digitization. It does not rely on massively annotated data sets or expensive model fine tuning, which is a huge plus for low- and middle-income countries. So, our resources are actually not only constrained by funds, but also by time, because MedPromptExtract helps us reduce the burden on emissions by automatically pulling out meaningful insights from unstructured text.

So that they can focus more on patient care instead of this paperwork. And because data is both anonymized and well structured, it becomes a valuable resource for downstream applications, research audits, or setting up signals, all the while protecting patient personal data.

Randye Kaye: Finally, the work focused on cases of acute kidney injury. So, looking forward, what are some potential other applications or expansions of MedPromptExtract?

Kshitij Jadhav: So looking ahead, there's a lot of potential to expand MedPromptExtract beyond just the cases of acute kidney injury. Like I said, we developed it as a precursor to our project on acute kidney injury, which has been also -- which

is close to completion also, but the architecture of MedPromptExtract is quite generalizable. Even though we tested it on the KI discharge summaries from a single hospital, the pipeline is actually very, very adaptable to be extended to extract data from other clinical conditions like cardiac events from ICU related studies or post-surgical complications. Depending on the use case, one will one will have to determine what data points have to be extracted of a particular medical condition. And in fact, by just -- think of it as a data, a disease centric use case in fact MedPromptExtract can also be used to extract data that may help streamline the hospital based administrative processes such as patient management related situations such as insurance claims and identification of high-risk patients for readmission risk mitigation.

So, with just some minor tweaks to the prompts and the training set up, the tool could be deployed in different scenarios in different settings, it is already up and running in a hospital setting, and the plan is to eventually integrate directly into the electronic health records, enabling real time data extraction and decision support at the point.

Randye Kaye: All right. Thank you so much. Very interesting. Thank you so much for joining us today.

Kshitij Jadhav: Thank you.

Randye Kaye: That was Dr. Kshitij Jadhav from the Artificial Intelligence and Digital Health Lab at Koita Centre for Digital Health of IIT Bombay, describing the *JALM* article, "MedPromptExtract (Medical Data Extraction Tool): Anonymization and High-Fidelity Automated Data Extraction Using Natural Language Processing and Prompt Engineering." Thanks for tuning into this episode of *JALM* Talk. See you next time, and don't forget to submit something for us to talk about.