Language Technology in Dermatology

Batorsky B, Mangialardi K

Our main purpose for this review is to see what lessons we can learn from the literature as we begin to build out a proof-of-concept LLM-based clinical summarization engine for structured dermatology reports. To that end, we want to understand the strengths and limitations of language technology for summarization of medical, specifically dermatological, information.

There were three main topics we covered in our review:

- Explore broadly how language technology is being used in dermatology
- Specifically understand how LLMs are being used in dermatology
- Outline the potentials and limitations of language technology for clinical summarization

Methodology

We conducted web searches along the lines of the research topics outlined above. We followed links and references from these sources to expand our coverage. This methodology should not be considered a systematic or comprehensive review, but a broad overview of the topics being discussed in the literature.

Artificial Intelligence (AI) applications in dermatology mainly focus on image-based diagnostics

In general, much of the focus of AI in dermatology has been on diagnostic use-cases. Our survey turned up several reviews of AI applications, but for the most part these applications were focused on image data and specifically on the use of AI systems to perform automated diagnosis.

There are a number of applications that have been released in the area of image-based diagnostics. For example First Derm provides a set of potential conditions based on an image input. However, the set of approved technologies lags behind that of other fields focused on diagnostic imaging like radiology. While radiology has a number of FDA approved applications, there are currently none for dermatology.

Our search pointed to several potential reasons for that:

1. Many diagnostic applications focus on binary objectives (present/not present), which does not reflect the variability in dermatological conditions and their diagnosis
2. Limited resolution on images taken by non-clinical devices (e.g. smartphones)
3. Diversity of skin tones leading to difficulty in model generalization and variability in performance

These issues are likely to be relevant in developing technology for clinical summarization in dermatology. Particularly, we should examine cases where the diagnoses are similar, but the presentation is different and understand whether and how a generative model struggles with these nuances.

Recently, generative AI has been used for diagnosis and prediction. SkinGPT combines image and text models to perform diagnosis, suggest treatments and simulate outcomes. The company behind the model, Haut.AI, has begun developing applications for the cosmetic and consumer markets.

Language technology has been applied in a variety of use-cases in dermatology

Language technology appears to have a long history with dermatology. In the 1980s, IBM developed an AI system called TEGUMENT / CLINIDERM to assist dermatopathological diagnosis based on descriptive inputs. This system 1) resolves all features of skin disease into several elementary, cutaneous forms, 2) considers descriptive input from physicians about the features of a specific case, and 3) provides a differential diagnosis. In the study cited above, pathologists made the same diagnosis as the program 92% of the time. It does not appear that the program ever gained widespread uptake among clinicians, but it does appear to be a foundational exploration of the potential of language technology in medical summarization in dermatology.

More commonly, language technology is used in the context of patient care and managing workflows. One example that is popular among dermatology clinics in the US is Klara. Klara is a HIPAA-compliant online care patient care platform that 1) coordinates communication among patients, providers, labs, and pharmacies, 2) streamlines administrative workflows, and 3) tracks patient data including imaging. The AI capabilities of the platform triage patient issues independently or route patients directly to the proper staff member who can address their concern. This is a commonly applied workflow using language technology in the clinical context.

One interesting application of language technology is in the DERMACLEAR study, which used natural language processing of structured and unstructured text from patient records to assess

prevalence of a set of dermatological conditions. The system achieved >95% precision, though their methodology for assessment was based on what was surfaced by the system, not pre-established ground truth. In addition to arriving at a diagnosis, the NLP tools were also utilized to assess correlations between patient demographics and lab tests and describe trends in treatment. Though the details of the tools they used are not clearly specified, this demonstrates an interesting approach to using language technology in the field.

Language technology shows promise in writing of clinical notes

Experiments with language technology using clinical data have a long history. As with computer vision technology, much of the focus has been on diagnostic applications or the identification of important clinical entities. For example, the clinical Text Analysis and Knowledge Extraction System (cTakes), originally developed in the late 2000s, uses a complex system of deterministic techniques to match free text to structured clinical vocabularies.

With the expanded summarization capabilities of language models comes the potential for additional applications to support the writing of clinical notes, a task that requires a considerable amount of a physician's time. Recent research has shown that LLMs are capable of writing comprehensive and accessible notes, though there is significant variability in the output depending on the input. For example, it suggested a different set of diagnoses depending on whether a bump was described as "red" versus "purple".

However, the field is evolving rapidly. A pre-print article this year showed that LLMs could outperform humans on several clinical tasks based on qualitative assessment. Many of the clinical tasks involve translating findings from radiology reports into impressions. Since parallels have been drawn between AI applications in radiology and dermatology, this may indicate these approaches have promise in the dermatology space as well. However, as mentioned above, variability in presentation may make this more complex. This paper also references several standard medical text datasets that may be useful for initial analysis.

The optimal solution appears to be having summarization systems include both AI and human experts. The AI is able to save the expert time while the expert is able to steer the AI and edit summarizes according to their analysis of the case. Augmenting, rather than replacing, human effort with AI has been widely recommended as the best approach to implementing AI systems.
.

For the purposes of summarizing structured dermatology reports, this means a fully autonomous system may be difficult to implement while ensuring consistent performance. Our strategy should focus on augmentation for the best results.

Conversational AI has been used with mixed success

Since ChatGPT came out, much of the focus in NLP has been on “conversational AI”; systems that generate responses conditional on the conversation context (i.e. the user input and system output). These have a wide array of applications, though they suffer from the same quality control issues that afflict LLMs more generally. In fact, these issues are compounded as the context (i.e. the conversation) expands.

In one study investigating the use of LLMs for medical education, researchers found that while LLMs offer a useful way to navigate large corpora of relevant text through “discussion” with the LLM, they were limited by the underlying data on which they are trained. Rare diseases tended to be overrepresented in text and, as a result, overrepresented in responses to queries from the LLMs.

Generally, conversational AI has been useful for creating chatbots that can help gather information and support scheduling. As described above, Klara is one company that is focused on these types of applications of conversational AI.

Conclusions

Dermatology is a field that has been seeing increasing usage of AI, but issues around diversity of data and the problem space itself has slowed adoption. Much of the focus of the field has been in the area of diagnosis based on images. However, as with all fields, the advent of advanced generative AI is pushing the boundaries of what is possible.

In the development of a summarization system, we need to be aware of the limitations that afflict them. General metrics such as BLEU scores may not accurately capture the quality of summaries, so they should be used alongside qualitative assessment. Inputs should be tweaked and the resulting changes in output should be evaluated. Further, inputs should certainly account for patient-specific factors that could contribute to differences in the presentation of a specific skin condition (e.g., Fitzpatrick skin type). Complex cases should be assessed against “simple” cases as there may be a bias in LLM models towards “rare” conditions. And performance should be assessed for individuals of different backgrounds to capture bias.

Conversational applications or applications involving long contexts may expose the limits of LLM systems. Input structure should be kept consistent to whatever extent possible to control the inherent variability in these models.

Though there is a great deal of potential in the use of these technologies, the clinical environment is a highly sensitive one and small errors can have major impacts on the course of treatment and the patients’ relationship with their clinician. The goal of a summarization technology should not be to “replace” the clinician’s work, but rather facilitate it. Based on

discussions with our physician colleagues, we've identified that much of the work in completing e-consultations is administrative (e.g. navigating the platform, completing the relevant fields, billing workflows). Metrics for performance for language technology systems should include time spent on consults. If writing clinical notes is a small portion of this time, the summarization engine would need to be extremely high quality to achieve significant savings.

In general, clinicians should be closely consulted to provide input on accuracy and utility of the output. Creating an additional workload would potentially reduce adoption and productivity.