Article

Not peer-reviewed version

One Step Closer to Conversational Medical Records: ChatGPT Parses Psoriasis Treatments from EMRs

Jonathan Shapiro *, Mor Atlas, Sharon Baum, Felix Pavlotsky, Aviv Barzilai, Rotem Gershon, Romi Gleicher, Itay Cohen

Posted Date: 29 September 2025

doi: 10.20944/preprints202509.2372.v1

Keywords: artificial intelligence; psoriasis; ChatGPT; electronic medical records

Article

One Step Closer to Conversational Medical Records: ChatGPT Parses Psoriasis Treatments from EMRs

Jonathan Shapiro [1,*], Mor Atlas [2], Sharon Baum [3,4], Felix Pavlotzky [3,4], Aviv Barzilai [3,4], Rotem Gershon [3,4], Romi Gleicher [5] and Itay Cohen [6]

[1] Maccabi Healthcare Services, Tel Aviv-Yafo, Israel
[2] Ono Academic College, Israel
[3] Dermatology Department, Sheba Medical Center, Ramat-Gan, Israel
[4] Gray Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel
[5] Ruth and Bruce Rappaport Faculty of Medicine, Technion - Institute of Technology, Israel
[6] Rutgers School of Public Health, Rutgers University, Piscataway, New Jersey, USA
* Correspondence: jonmidi@gmail.com

Abstract

Large Language Models (LLMs), such as ChatGPT, are increasingly used in clinical settings, including documentation and decision support. However, their accuracy in extracting treatment data from unstructured dermatology records remains underexplored. We evaluated ChatGPT-4o's ability to identify psoriasis treatments from free-text documentation, compared with expert annotations. Ninety-four electronic medical records (EMRs) of psoriasis patients were retrospectively analyzed. ChatGPT-4o extracted treatments from each note, and its output was compared to manually curated annotations by dermatologists. Eighty-three treatments, including topical agents, systemic medications, biologics, phototherapy, and procedures, were evaluated. Performance metrics included recall, precision, F1-score, specificity, accuracy, Cohen's Kappa, and AUC. ChatGPT-4o demonstrated strong performance, with a recall of 0.91, a precision of 0.96, an F1-score of 0.94, a specificity of 0.99, and an accuracy of 0.99. Agreement with expert annotations was high (Cohen's Kappa = 0.93; AUC = 0.98). Group-level analysis confirmed these results, with the highest performance in biologics and methotrexate (F1 = 1.00) and lower recall in categories with vague documentation, such as systemic steroids and antihistamines. Our study highlights the potential of LLMs to extract psoriasis treatment information from unstructured clinical documentation and structure it for research and decision support. The model performed best with well-defined, commonly used treatments.

Keywords: artificial intelligence; psoriasis; ChatGPT; electronic medical records

Introduction

The integration of large language models (LLMs), such as ChatGPT, into clinical informatics has opened promising avenues in medical documentation, decision support, and research. LLMs have shown capabilities in automating administrative tasks, summarizing patient information, and enhancing clinician workflows and patient communication [1,2]. However, their application in dermatology, particularly in parsing unstructured clinical narratives to extract treatment histories, remains underexplored.

Psoriasis is a chronic, immune-mediated skin disease affecting approximately 3% of the global population. Its management involves individualized, often complex regimens, including topical therapies, systemic medications, biologics, phototherapy, and procedural interventions [2,3]. Given the disease's unpredictable course, risk of progression to psoriatic arthritis, and variability in treatment response, tracking longitudinal treatment history is both clinically essential and operationally challenging. While this information is often documented in electronic medical records

 doi:10.20944/preprints202509.2372.v1

(EMRs), it typically resides in free-text form, which lacks standardization and is challenging to analyze retrospectively [3,4].

Recent advances in natural language processing (NLP) have enabled AI-powered tools to enhance electronic medical records (EMR) abstraction, with studies showing improved disease identification when narrative text is analyzed alongside structured codes, outperforming code-only algorithms in diseases like psoriatic arthritis and inflammatory dermatoses [3,5]. In dermatology, tools like ChatGPT have demonstrated potential in patient education, clinical summarization, and simulation of board exam scenarios [1,6]. Specifically in psoriasis, ChatGPT-4 has been used to identify affected body areas and comorbidities from clinical narratives [7], with some success in comparing treatment options and supporting patient engagement, although diagnostic limitations remain [2,5]. In rheumatology, NLP-based classification of psoriatic arthritis has outperformed rule-based models, further supporting the feasibility of LLMs in treatment-related tasks [3].

Despite these advances, the capacity of LLMs, particularly ChatGPT, to extract and structure detailed treatment histories from unstructured EMRs in psoriasis remains underinvestigated. Clinical documentation often includes variable terminology, abbreviations, spelling inconsistencies, and linguistic ambiguity. Moreover, distinguishing between treatments for psoriasis and those for comorbidities adds further complexity [8,9]. Phrases involving negation or uncertainty (e.g., "the patient was not treated with methotrexate") must also be interpreted correctly to avoid false positives [10,11].

LLMs like GPT-4 have shown promise in overcoming similar challenges in other domains. For example, they have demonstrated high recall in extracting findings from radiology reports (99.3%), supporting the feasibility of automating data extraction tasks[12]. They have also been effective in de-identifying clinical notes and generating synthetic data, thereby improving administrative workflows and data quality [9,13]. In oncology, ChatGPT-4 has been used to identify cancer phenotypes from EHR text, further supporting its value in extracting clinical information from unstructured narratives[14]. Moreover, ChatGPT-4 has achieved diagnostic accuracy levels comparable to physicians when identifying final diagnoses from differential lists (footnote 1).

These findings suggest a strong potential for LLMs to support automated extraction of structured treatment data from complex, unstructured text, which is a critical need in dermatology, where documentation practices are exceptionally heterogeneous.

In this exploratory study, we evaluate the general-purpose language model ChatGPT-4o's ability to extract treatment information from unstructured EMRs of patients with psoriasis. By comparing its outputs to gold-standard annotations from expert dermatologists, we assess its accuracy and explore its potential utility in dermatology workflows. This work contributes to the growing body of literature supporting the use of LLMs in dermatology. It aims to inform scalable AI-assisted solutions for clinical documentation, retrospective cohort assembly, and decision support in chronic dermatoses.

Materials and Methods

We retrospectively reviewed 94 electronic medical records (EMRs) of patients diagnosed with psoriasis and treated at the Dermatology and Psoriasis Clinic at Sheba Medical Center. The EMRs were written in a hybrid of Hebrew and English and included patient anamnesis, physical examination findings, treatment history, and management plans. A board-certified dermatologist manually reviewed each record to annotate all treatments specifically administered for psoriasis. Treatments documented for other dermatologic conditions were ignored. Each patient case was transformed into a structured dataset entry, with binary indicator variables corresponding to 83 possible treatments. Each treatment was labeled TRUE if used for treating psoriasis and FALSE otherwise, resulting in a sparse matrix.

To evaluate automated extraction, we applied ChatGPT-4o, a general-purpose multimodal language model developed by OpenAI. A custom GPT agent was created on OpenAI's platform (psoriasis-treatment-extractor), with the following exact instructions: *'The GPT will read a mix of*

 doi:10.20944/preprints202509.2372.v1

Hebrew and English summaries provided by the user, extracting only the treatments related to psoriasis. If a treatment is mentioned, the GPT will identify whether the patient received it, note the patient's response to the treatment, and specify the duration or number of treatment courses if mentioned. The GPT will avoid mentioning treatments for other diseases and focus solely on psoriasis-related information. Mention only treatments in the past and not treatments planned for the future.' The model was not provided with a predefined medication list and relied solely on the textual content of each EMR. Each note was entered individually in a new session, with the model instructed not to save any data for future training. The extracted treatments were then manually copied into a binary-coded table identical in structure to the expert-annotated dataset.

We then obtained parallel binary classifications—one from the human expert and one from ChatGPT-4o—for each of the 94 records and 83 treatments. To assess model performance, we calculated accuracy, precision, recall, F1-score, specificity, Cohen's Kappa, and the area under the receiver operating characteristic curve (AUC).

Evaluation was conducted at multiple levels: (1) a global level, treating each treatment-patient combination as a separate binary instance (7,802 in total); (2) treatment-wise, where each treatment was evaluated as a separate classification task (limited to treatments with at least five positive cases); and (3) group-level, where treatments were clustered into pharmacologic categories (e.g., biologics, topical steroids) to increase statistical stability and interpretability. Table 1 outlines the pharmacologic categories and their corresponding treatments used to group medications for the group-level analysis. Lastly, we computed per-record Cohen's Kappa to assess agreement on a patient-level basis.

Table 1. Pharmacologic Categories and Included Treatments for Psoriasis.

Pharmacologic Group	Included Treatments
Topical Steroids (TCS)	Dermovate, Betacorten G, Diprosalic, Diprospan, Topical steroid ointments, etc.
Topical Vitamin D Analogues	Daivobet, Xamiol
Topical Steroid + Vitamin D Analogues	Daivobet, Xamiol (combination products)
Topical Steroid + Salicylic Acid	Diprosalic, Topisalen
Salicylic Acid	Topisalen, Salicylic preparations
TAR-based Topicals	Goeckerman protocol, TAR ointments
Systemic Retinoids	Neotigasone
Methotrexate (MTX)	MTX
Biologic Therapy	Humira, other biologics not individually named
Systemic Steroids	Prednisone, Diprospan
Steroids – Intralesional (IL)	Betacorten G, IL steroids
Phototherapy	Phototherapy (UV), Goeckerman protocol
Antihistamines	Loratadine, others not explicitly named
Systemic Antibiotics	Augmentin PO, others not explicitly named
Moisturizers	Emollients, unspecified moisturizers

All analyses were conducted using R (version 4.4.1).

IRB approval status: Reviewed and approved by the Sheba Medical Center Ethics Committee, approval number 1083-24-SMC.

Results

Of the 94 psoriasis cases included in the study, 55 were female (58.5%) and 39 were male (41.5%). The age range of patients was 18.9 to 86.7 years. Clinical visit notes averaged 278 ± 154 words in length. The number of psoriasis treatments per patient varied from zero (in five cases) to twelve, with a median of three and a mean of 3.2. Some treatments, including Loratadine, Dexacort, and IV

 doi:10.20944/preprints202509.2372.v1

Experimental therapy, were rare—appearing only once—and were missed by the model. Conversely, ChatGPT-4o misidentified three treatments as psoriasis-related that were actually indicated for other conditions.

At the global level, ChatGPT-4o demonstrated strong overall performance across 7,802 binary instances (83 treatments × 94 records). It achieved a recall of 0.91 and a precision of 0.96, resulting in an F1-score of 0.94. Specificity and accuracy both reached 0.99, and Cohen's Kappa was 0.93, indicating strong agreement with expert annotations beyond chance. The model's AUC was 0.98, reflecting excellent discrimination between prescribed and non-prescribed treatments.

A treatment-wise evaluation, limited to treatments administered in at least five patient records, confirmed these trends. Perfect precision (1.00) was observed for nearly all treatments, indicating a low false positive rate. Recall values were more variable, ranging from 0.77 for treatments such as unspecified topical ointments to 1.00 for well-documented therapies including Humira, MTX, Phototherapy, and Dermovate. The F1-score exceeded 0.90 for most treatments. Specificity was consistently near 1.00. Cohen's Kappa values ranged from 0.79 to 1.00, with most above 0.86, and AUC scores were uniformly strong, all exceeding 0.88. Table 2 presents the full set of performance metrics for these treatments, while Figure 1 illustrates their respective recall values.

Table 2. Per-treatment performance metrics for treatments with ≥5 positive Instances.

Treatment	Precision	Recall	F1	SP	AU	AC	KP	TP	TN	FP	FN	Positive cases
Humira	1	1	1	1	1	1	1	6	88	0	0	6
Goeckerman protocol	1	0.83 ± 0.30	0.91 ± 0.23	1	0.92	0.99± 0.02	0.9	5	88	0	1	6
MTX	1	1	1	1	1	1	1	17	77	0	0	17
Phototherapy	1	1	1	1	1	1	1	29	65	0	0	29
Diprosalic	1	1	1	1	1	1	1	6	88	0	0	6
Topisalen	1	0.86 ± 0.26	0.92 ± 0.20	1	0.93	0.99± 0.02	0.92	6	87	0	1	7
Xamiol	1	1	1	1	1	1	1	7	87	0	0	7
Diprospan	1	1	1	1	1	1	1	7	87	0	0	7
Prednisone	1	0.78 ± 0.27	0.88 ± 0.22	1	0.89	0.98± 0.03	0.86	7	85	0	2	9
Betacorten G	1	1	1	1	1	1	1	10	84	0	0	10
Dermovate	0.88 ± 0.15	1	0.94 ± 0.12	0.97 ± 0.03	0.99	0.98± 0.03	0.92	15	77	2	0	15
Steroid ointments	1	0.82 ± 0.18	0.90 ± 0.14	1	0.91	0.97± 0.04	0.88	14	77	0	3	17
Topical ointments	0.96 ± 0.08	0.77 ± 0.15	0.85 ± 0.13	0.98 ± 0.03	0.88	0.91± 0.06	0.79	23	63	1	7	30
Neotigasone	1	1	1	1	1	1	1	12	82	0	0	12
Daivobet	0.80 ± 0.35	0.80 ± 0.35	0.80 ± 0.35	0.99 ± 0.02	0.89	0.98± 0.03	0.79	4	88	1	1	5

SP – Specificity, AU – Area Under the Curve, AC – Accuracy, KP = Kappa, TP – True Positive, TN – True Negative, FP – False Positive, FN – False Negative.

 doi:10.20944/preprints202509.2372.v1

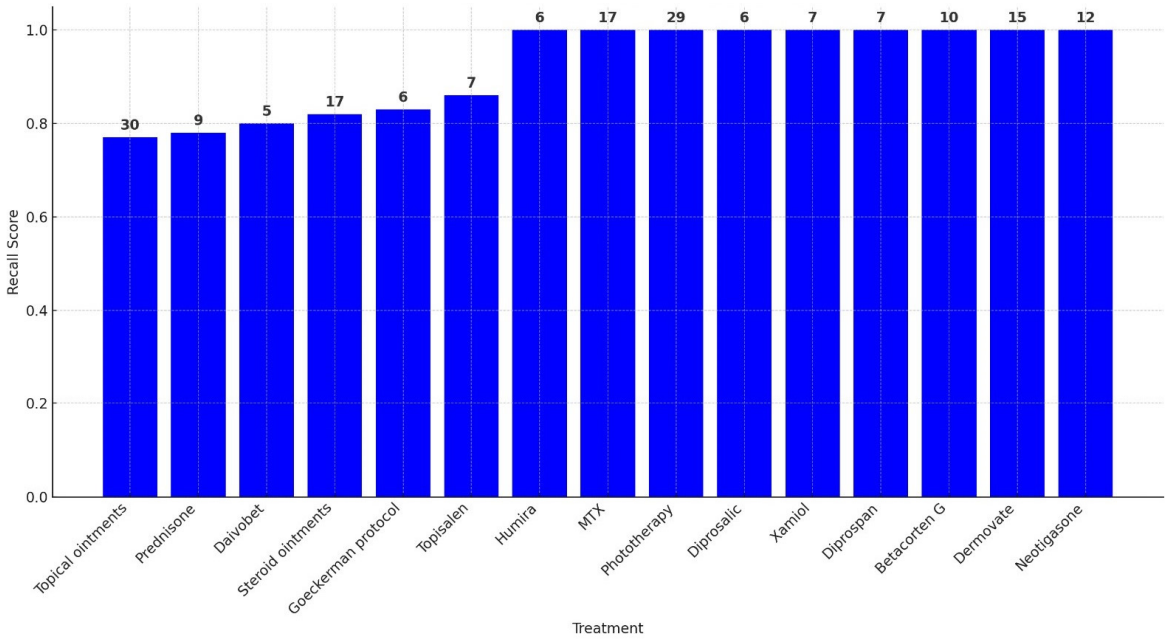

Figure 1. Recall score of treatment with more than five positive cases. The number on top of each column represents the number of positive cases in the database.

To enhance statistical stability, a group-level analysis was performed by aggregating treatments into pharmacologic categories. We included 15 of 23 possible categories, excluding those with fewer than five total observations. Categories included biologics, topical steroids, systemic antibiotics, antihistamines, and systemic retinoids. ChatGPT-4o maintained excellent precision across all included groups, ranging from 0.85 to 1.00. Recall values ranged from 0.70 (e.g., systemic steroids, antihistamines) to 1.00 (e.g., biologics, MTX, salicylic acid), with most groups exceeding 0.89. Group-level F1-scores were generally ≥0.90, and specificity and accuracy remained high. AUC values consistently exceeded 0.92 across all categories. Cohen's Kappa was above 0.90 in nearly all groups. Table 3 presents the complete set of performance metrics for each treatment category. Figure 2 provides a visual comparison of precision and recall values across these categories. Despite variability in recall for a few categories, particularly systemic steroids and antihistamines, the model's precision and agreement with expert annotation remained consistently high. These findings underscore ChatGPT-4o's strong performance in extracting real-world psoriasis treatments across diverse pharmacologic domains and documentation styles.

Table 3. Group-level Treatment Performance Metrics for groups with more than five positive cases. Medication represents the number of treatments in the group.

Group	MD	N	Pre	Recall	F1	SP	AUC	AC	Kappa	TP	TN	FP	FN
Antibiotics - systemic	8	18	1	0.89 ± 0.15	0.94± 0.11	1	0.94	1.00 ± 0.00	0.94	16	734	0	2
Anti - histamines	9	17	0.92 ± 0.14	0.71 ± 0.22	0.80± 0.19	1.00± 0.00	0.85	0.99 ± 0.01	0.8	12	828	1	5
Biologic therapy	7	20	1	1	1	1	1.00	1	1	20	638	0	0
Goeckerman	1	6	1	0.83 ± 0.30	0.91± 0.23	1	0.92	0.99 ± 0.02	0.9	5	88	0	1
Moisturizer	6	12	0.85 ± 0.20	0.92 ± 0.16	0.88± 0.18	1.00± 0.01	0.96	0.99 ± 0.01	0.88	11	550	2	1
MTX	1	17	1	1	1	1	1.00	1	1	17	77	0	0
Phototherapy	2	33	1	0.97 ±	0.98± 0.04	1	0.98	0.99 ± 0.01	0.98	32	155	0	1

 doi:10.20944/preprints202509.2372.v1

				0.06									
Salicylic acid	2	6	0.86 ± 0.26	1	0.92± 0.21	0.99± 0.01	1.00	0.99 ± 0.01	0.92	6	181	1	0
Topical Steroid + Salicylic acid	2	13	1	0.92 ± 0.14	0.96± 0.11	1	0.96	0.99 ± 0.01	0.96	12	175	0	1
Topical Steroid + vitamin D	2	12	0.92 ± 0.16	0.92 ± 0.16	0.92± 0.16	0.99± 0.01	0.96	0.99 ± 0.01	0.91	11	175	1	1
Steroids – Intralesional (IL)	1	7	1	1	1	1	1.00	1	1	7	87	0	0
Steroids - systemic	2	10	1	0.70 ± 0.28	0.82± 0.24	1	0.85	0.98 ± 0.02	0.82	7	178	0	3
Systemic retinoids	2	13	1	1	1	1	1.00	1	1	13	175	0	0
TAR - topical	3	8	0.89± 0.21	1	0.94± 0.16	1.00± 0.01	1.00	1.00 ± 0.01	0.94	8	273	1	0
TCS (Topical steroids)	12	85	0.96± 0.04	0.88 ± 0.07	0.92± 0.06	1.00± 0.00	0.94	0.99 ± 0.01	0.91	75	1040	3	10

MD - Medication, N= Number of cases, Pre - Precision, SP – Specificity, AUC - Area Under the Curve, AC – Accuracy, TP – True Positive. TN - True Negative, FP – False Positive, FN – False Negative.

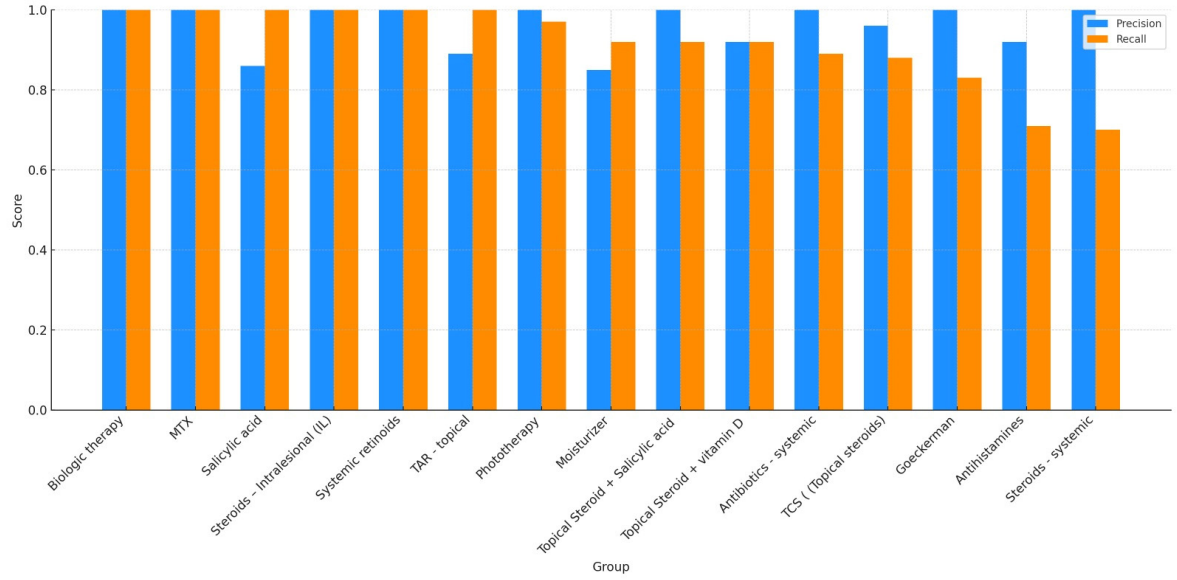

Figure 2. Group-level precision and recall.

Discussion

The current era of healthcare is marked by unprecedented growth in clinical data, both in volume and in complexity. Electronic medical records (EMRs), originally introduced to streamline documentation and improve continuity of care, have paradoxically created new burdens on clinicians. Numerous studies have highlighted that physicians now spend more time entering data into EMRs than they do in face-to-face patient encounters, with documentation often extending beyond clinical hours and contributing substantially to professional burnout [17,18]. This phenomenon, sometimes referred to as “pajama time,” highlights how the promise of digitization has not yet fully materialized in everyday practice. Instead of improving efficiency, EMRs have frequently added layers of administrative work that detract from clinical care.

The development and integration of artificial intelligence (AI) systems capable of parsing unstructured medical text and generating structured visit summaries is not merely a technological innovation but a systemic necessity. Healthcare systems worldwide are grappling with rising patient volumes, workforce shortages, and escalating costs. These pressures magnify the need for scalable solutions that can alleviate the administrative load while simultaneously improving the fidelity of

 doi:10.20944/preprints202509.2372.v1

medical data. In this context, large language models (LLMs) such as ChatGPT represent a paradigm shift: rather than asking clinicians to adapt their documentation practices to rigid data-entry forms, AI systems can adapt to the natural language of clinicians, extracting key details with high precision and recall [19].

Furthermore, the transition from free-text narratives to structured, analyzable data is a cornerstone for modern healthcare priorities such as value-based care, population health management, and precision medicine. Longitudinal tracking of treatments, responses, and adverse events is indispensable for both individual patient care and research into real-world treatment outcomes. Historically, this has required labor-intensive chart review and manual abstraction, limiting the scale and timeliness of insights. By automating the structuring of treatment histories directly from narrative notes, AI-based systems can unlock vast amounts of previously inaccessible information, enabling rapid cohort identification, large-scale retrospective analyses, and data-driven decision support [20].

It is important to situate this transformation within a broader technological trajectory. The past decade has witnessed incremental advances in natural language processing applied to clinical documentation, ranging from rule-based systems to statistical models and domain-specific machine learning approaches. Each stage improved performance but remained constrained by brittle rules or narrow vocabularies. The emergence of general-purpose LLMs, trained in vast multilingual corpora and capable of contextual reasoning, represents a step change in capability. Unlike earlier systems, these models are not limited to recognizing predefined keywords or phrases but can interpret ambiguous descriptions, handle linguistic variability, and infer clinical relevance across diverse contexts. This allows them to engage with EMRs in a manner that is far closer to human clinical reasoning, bridging the gap between free-text notes and structured data repositories [21].

Beyond efficiency and accuracy, the integration of AI into clinical documentation has broader implications for the physician–patient relationship and the future of medical practice. By reducing the time spent on clerical tasks, AI-assisted systems have the potential to restore clinicians' focus to direct patient care, improving satisfaction on both sides of the clinical encounter [22]. At the same time, the ability to generate comprehensive, accurate, and standardized summaries enhances communication among multidisciplinary teams, reduces the risk of missed information, and supports continuity of care across healthcare settings.

From a research standpoint, these developments also herald a new era in medical knowledge generation. The ability to query large EMR databases with AI tools enables researchers to identify patient subgroups, treatment patterns, and outcome trajectories at a speed and scale that manual chart review could never achieve. This creates opportunities for rapid hypothesis generation, real-world evidence studies, and post-marketing surveillance, all of which are increasingly recognized as essential complements to randomized controlled trials in guiding clinical decision-making.

The adoption of advanced AI tools such as ChatGPT-4o for treatment extraction from EMRs offers significant advantages for both clinical practice and research. In the clinical setting, one of the most promising benefits is the potential to generate comprehensive summaries of a patient's treatment history with reduced effort and improved accuracy. Traditionally, compiling such summaries during patient intake or follow-up visits requires clinicians to sift through lengthy and often fragmented medical records manually, a process that is both time-consuming and prone to error. While the current use of AI-powered extraction does not yet result in substantial time savings, it introduces a structured and consistent approach to retrieving relevant treatments, which may reduce the risk of missing critical information[9,15]. As technology continues to evolve, it may also lead to meaningful time savings in clinical workflows. Moreover, once generative AI models can consistently extract treatments with high precision, EMRs can serve as contextual input for interactive models, enabling the use of clinical notes as file-based data sources in a Retrieval-Augmented Generation (RAG) framework, forming the basis for informed, case-specific conversations with an AI assistant.

 doi:10.20944/preprints202509.2372.v1

For research, the implications are equally profound. AI models can be leveraged to perform large-scale cohort identification and retrospective analyses that would be infeasible with manual review. For instance, researchers can efficiently query the EMR database to identify all patients who did not respond to methotrexate (MTX) or who have been treated with a specific medication, supporting studies of treatment effectiveness, safety, and real-world outcomes[9,14]. This capability accelerates the pace of clinical research, enables rapid hypothesis generation and testing, and facilitates the development of precision medicine approaches by uncovering nuanced patterns in treatment response across diverse patient populations. As these tools continue to evolve, their integration into clinical and research workflows promises to enhance the quality, efficiency, and impact of both patient care and medical discovery.

This study explored ChatGPT-4o's ability to accurately identify therapies associated with psoriasis from complex, multilingual (Hebrew-English hybrid) medical records, distinguish them from treatments for other conditions, and compare its extraction accuracy to that of expert human annotation. The goal determined the feasibility and reliability of integrating advanced AI tools into clinical and research workflows for the automated identification of treatments in dermatology.

There is an abundance of LLMs available in the AI ecosystem. We chose OpenAI's GPT because it has the highest market share (footnotes 2,3) and is widely used as a research baseline (footnote 4). In addition, this study builds on our previous work with ChatGPT[7]. There are also open-sourced models that can be used, and even further fine-tuned to a specific domain, such as Mistral[16]. However, in this work, we aimed to evaluate the performance of a general-purpose model in informing and guiding potential users, focusing on the most commonly used publicly accessible option. It is worth noting that other models may achieve higher accuracy, particularly when trained explicitly on EMRs.

One main challenge in extracting treatment histories from unstructured EMRs is the variability in how treatments are described. In some cases, the documentation includes specific drug names, such as methotrexate or calcipotriol, which the model can more reliably recognize as administered treatments. However, in other instances, the clinical notes refer to general treatment categories, such as "topical treatments," "topical steroids," or "biologic therapies", without naming a particular medication. While these general terms likely indicate actual therapeutic use, they pose a challenge for the model in determining whether and how to classify them as concrete treatments for psoriasis.

In our study, ChatGPT-4o demonstrated high overall performance, with an accuracy of 0.99, a precision of 0.96, a recall of 0.91, and an F1 score of 0.94. Cohen's Kappa (0.93) and AUC (0.98) further supported its excellent discriminative ability and agreement with expert annotations. Treatment-specific and pharmacologic-group analyses highlighted strong overall metrics, although some variability in recall was observed, particularly for treatments with less explicit or more ambiguous documentation.

While ChatGPT-4o performed robustly in treatment identification, a deeper exploration of groups with lower recall provides important insights into model behavior and highlights areas for refinement. For example, the relatively low recall for systemic steroids and antihistamines, despite their presence in many patient records, corresponds with the lower Cohen's Kappa values for these treatment groups listed in Table 3 (antihistamines $\kappa = 0.79$; systemic steroids $\kappa = 0.77$). These findings likely reflect not only model limitations but also clinical and contextual ambiguity. Systemic steroids are generally not recommended for psoriasis due to the risk of rebound flares, and their use is typically confined to atypical cases or early diagnostic workups. In several instances, patients may have received systemic steroids before a definitive diagnosis of psoriasis was established, making it technically correct to list them as part of the treatment history, but potentially confusing for the model, which lacks the ability to infer temporal context or diagnostic intent. Similarly, antihistamines are often prescribed to patients with psoriasis suffering from severe pruritus, yet they are not disease-modifying agents and may be recorded as general symptomatic treatments. This variability in clinical context, coupled with nonspecific phrasing in electronic medical records (EMRs), likely contributed to the model's reduced ability to identify these therapies as psoriasis-related treatments consistently.

 doi:10.20944/preprints202509.2372.v1

In other instances, ChatGPT-4o identified medications that the human investigator did not explicitly mention.

A detailed examination of these lower-performing cases revealed several distinct patterns: One notable source of error was ChatGPT-4o incorrectly interpreting planned future treatments as treatments already administered. This misunderstanding primarily stemmed from ambiguities in clinical documentation, where future treatment plans were not clearly distinguished from historical or ongoing treatments. While this distinction is clear to human clinicians due to context or clinical familiarity, ChatGPT-4o occasionally struggled with this nuance. This emphasizes the need for more explicit differentiation within clinical notes between historical, current, and future treatment plans to enhance NLP accuracy. However, as the models improve, we believe their ability to distinguish past treatments from recommendations will also improve.

An interesting and proactive behavior exhibited by ChatGPT-4o was its holistic approach to psoriasis patient management. ChatGPT-4o occasionally suggested medications not specifically annotated by the human reviewer as psoriasis treatments but identified them as potential therapies due to their management of known comorbidities associated with psoriasis. For instance, ChatGPT-4o spontaneously listed statins used for treating hypercholesterolemia as potentially relevant treatments, justifying their inclusion by citing the well-established association between psoriasis and metabolic syndrome. Similarly, ChatGPT-4o independently considered psychiatric medications prescribed for depression as related to psoriasis management, due to the documented association between psoriasis and depression.

Furthermore, ChatGPT-4o proactively identified in some cases antifungal treatments prescribed empirically for suspected tinea pedis, reasoning that these treatments could be relevant in cases where clinicians might face diagnostic ambiguity between psoriasis and fungal infections or when treating fungal infections complicating psoriasis plaques. GPT-4o provided these additional interpretations with explanatory remarks, allowing human reviewers to explicitly determine their relevance.

Other medications proactively listed by ChatGPT-4o included liraglutide (Saxenda) for obesity management, reflecting careful consideration of obesity as an important comorbidity that dermatologists closely monitor in patients with psoriasis. ChatGPT-4o also noted medical cannabis usage, expressing appropriate reservations about its explicit classification as a psoriasis treatment but nevertheless flagging its potential clinical relevance.

These proactive inclusions by ChatGPT-4o highlight the sophisticated, clinically informed reasoning that AI models can perform, surpassing the strict labeling provided by human reviewers. However, this also suggests the necessity of clear instructions in AI prompts regarding whether to include or explicitly exclude treatments for psoriasis-associated comorbidities, depending on the intended clinical or research context. For research purposes, this ability may lead to more insights into treatment options and patterns.

Another practical reason for treatment misidentification observed was ChatGPT-4o's occasional misreading or incorrect interpretation of medication names. For instance, ChatGPT-4 mistakenly read "Dermovate" instead of "Daivobet" (both spelled in Hebrew), a misinterpretation that highlights the inherent limitations of AI in recognizing medication names. This demonstrates that, despite overall impressive accuracy, the model can still make specific identification errors, reinforcing the continued necessity of human oversight. These types of errors are also expected to decrease with more training and model improvements.

This work should be viewed within the broader transformation taking place in medicine. Across specialties, AI tools are being developed to analyze patient records and generate structured summaries for both clinical practice and research. Our study demonstrates how this vision can be operationalized: using a general-purpose language model, we showed that treatment information embedded in dermatology notes can be reliably extracted and organized. Although psoriasis was chosen as an initial test case, the approach extends to other chronic diseases where treatment histories are equally complex and clinically significant. By demonstrating that a non-specialized LLM can

 doi:10.20944/preprints202509.2372.v1

achieve high accuracy in parsing treatments from routine EMRs, this work provides proof of concept for scalable AI-assisted medical record summarization. Such capabilities have the potential to reduce clinicians' documentation burden, strengthen continuity of care, and unlock new opportunities for real-world data analysis across the medical field.

Limitations

Our study had several limitations. First, due to the limited number of positive instances for several treatment labels, the statistical analysis is primarily descriptive. Future research with a substantially larger dataset is needed to more reliably assess the model's sensitivity in identifying different treatments. Second, ChatGPT-4o was not pre-trained on a predefined list of psoriasis treatments. Although overall performance was high, we expect that results would further improve with fine-tuning on domain-specific treatment vocabularies. Third, the unstructured EMR data consisted of a mix of Hebrew and English. We assume that performance on fully English-based EMRs would likely be higher. Another limitation of this study is the relatively modest sample size of 94 records. Although the dataset generated more than 7,800 binary treatment instances, many treatments were represented only a handful of times, which constrained statistical power and reduced the stability of per-treatment performance metrics. To validate these findings and improve robustness, larger multi-center datasets will be necessary, enabling greater generalizability across diverse populations and documentation practices. Future research may also explore temporal parsing of treatment sequences or the use of fine-tuned language models trained specifically on dermatologic corpora to enhance context recognition and extraction accuracy.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was approved by the Sheba Medical Center Institutional Review Board (IRB), approval number 1083-24-SMC. The study was not registered in a clinical trials registry as it does not involve an interventional clinical trial.

Informed Consent Statement: As this was a retrospective analysis of anonymized medical records, no informed consent was required.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon request, owing to privacy/ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest. No funding was received for this study. The authors used ChatGPT-4o (OpenAI) during manuscript preparation to assist with language and editing. All content was reviewed and approved by the authors, who take full responsibility for the final version.

Acknowledgments: We would like to express our sincere gratitude to Prof. H. Peter Soyer, Chair in Dermatology at the University of Queensland's Frazer Institute, for his expert guidance and valuable support.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the author used ChatGPT-4o to improve language and readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Footnotes – (Non-Peer-Reviewed Sources)

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. *Capabilities of GPT-4 on Medical Challenge Problems*. arXiv preprint. 2023. Available from: https://arxiv.org/abs/2303.13375
2. PR Newswire. *New Statcounter AI Data Finds ChatGPT Sends 79.8% of All Chatbot Referrals to Websites*. 2025. Available from: https://www.prnewswire.com/news-releases/new-statcounter-ai-data-finds-chatgpt-sends-798-of-all-chatbot-referrals-to-websites-301965483.html
3. Statcounter. *AI Chatbot Market Share Worldwide*. Statcounter Global Stats. 2025. Available from: https://gs.statcounter.com/ai-chatbot-market-share

 doi:10.20944/preprints202509.2372.v1

4. Xu Z. *Patterns and Purposes: A Cross-Journal Analysis of AI Tool Usage in Academic Writing*. arXiv preprint. 2025. Available from: https://arxiv.org/abs/2502.00632

References

1. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: Practical implications within dermatology. J Am Acad Dermatol. 2023 Oct;89(4):870-871. doi: 10.1016/j.jaad.2023.05.081. Epub 2023 Jun 12. PMID: 37315798.
2. Ravipati A, Elman SA. The state of artificial intelligence for systemic dermatoses: Background and applications for psoriasis, systemic sclerosis, and much more. Clin Dermatol. 2024 Sep-Oct;42(5):487-491. doi: 10.1016/j.clindermatol.2024.06.019. Epub 2024 Jun 21. PMID: 38909858.
3. Love TJ, Cai T, Karlson EW. Validation of psoriatic arthritis diagnoses in electronic medical records using natural language processing. Semin Arthritis Rheum. 2011 Apr;40(5):413-20. doi: 10.1016/j.semarthrit.2010.05.002. Epub 2010 Aug 10. PMID: 20701955; PMCID: PMC3691811.
4. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016 Sep;23(5):1007-15. doi: 10.1093/jamia/ocv180. Epub 2016 Feb 5. PMID: 26911811; PMCID: PMC4997034.
5. Perrin J, Petronic-Rosic V. The potential role and restrictions of artificial intelligence in medical school dermatology education. Clin Dermatol. 2024 Sep-Oct;42(5):477-479. doi: 10.1016/j.clindermatol.2024.06.017. Epub 2024 Jun 24. PMID: 38925446.
6. Goktas P, Grzybowski A. Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review. J Clin Med. 2024 Oct 3;13(19):5909. doi: 10.3390/jcm13195909. PMID: 39407969; PMCID: PMC11477344.
7. Shapiro J, Baum S, Pavlotzky F, Mordehai YB, Barzilai A, Freud T, Gershon R. Application of a natural language processing artificial intelligence tool in psoriasis: A cross-sectional comparative study on identifying affected areas in patients' data. Clin Dermatol. 2024 Sep-Oct;42(5):480-486. doi: 10.1016/j.clindermatol.2024.06.018. Epub 2024 Jun 21. PMID: 38909857.
8. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79. PMID: 18660879; PMCID: PMC2592252.
9. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H. Clinical information extraction applications: A literature review. J Biomed Inform. 2018 Jan;77:34-49. doi: 10.1016/j.jbi.2017.11.011. Epub 2017 Nov 21. PMID: 29162496; PMCID: PMC5771858.
10. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10. doi: 10.1006/jbin.2001.1029. PMID: 12123149.
11. Sohn S, Wagholikar KB, Li D, Jonnalagadda SR, Tao C, Komandur Elayavilli R, Liu H. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):836-42. doi: 10.1136/amiajnl-2013-001622. Epub 2013 Apr 4. PMID: 23558168; PMCID: PMC3756269.
12. Hasani AM, Singh S, Zahergivar A, Ryan B, Nethala D, Bravomontenegro G, Mendhiratta N, Ball M, Farhadi F, Malayeri A. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol. 2024 Jun;34(6):3566-3574. doi: 10.1007/s00330-023-10384-x. Epub 2023 Nov 8. PMID: 37938381.
13. Altalla' B, Abdalla S, Altamimi A, Bitar L, Al Omari A, Kardan R, Sultan I. Evaluating GPT models for clinical note de-identification. Sci Rep. 2025 Jan 31;15(1):3852. doi: 10.1038/s41598-025-86890-3. PMID: 39890969; PMCID: PMC11785955.
14. Bhattarai K, Oh IY, Sierra JM, Tang J, Payne PRO, Abrams Z, Lai AM. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. JAMIA Open. 2024 Jul 3;7(3):ooae060. doi: 10.1093/jamiaopen/ooae060. PMID: 38962662; PMCID: PMC11221943.
15. Dobry A, Begaj T, Mengistu K, Sinha S, Droms R, Dunlap R, Wu D, Adhami K, Stavert R. Implementation and Impact of a Store-and-Forward Teledermatology Platform in an Urban Academic Safety-Net Health

 doi:10.20944/preprints202509.2372.v1

Care System. Telemed J E Health. 2021 Mar;27(3):308-315. doi: 10.1089/tmj.2020.0069. Epub 2020 Jun 9. PMID: 32522105.

16. A.R. Randhawa, A.; Zakka, C.; Hiesinger, W.; Sattar, M., Comparative Analysis Of Language Model Systems In Endocrinology: Performance And Human Acceptability Assessment, Journal of the Endocrine Society, 8 (2024) bvae163.1038. https://doi.org/10.1210/jendso/bvae163.1038
17. Holmgren AJ, Apathy NC, Sinsky CA, Adler-Milstein J, Bates DW, Rotenstein L. Trends in Physician Electronic Health Record Time and Message Volume. JAMA Intern Med. 2025 Apr 1;185(4):461-463. doi: 10.1001/jamainternmed.2024.8138. PMID: 39992635; PMCID: PMC11851296.
18. Tajirian T, Lo B, Strudwick G, Tasca A, Kendell E, Poynter B, Kumar S, Chang PB, Kung C, Schachter D, Zai G, Kiang M, Hoppe T, Ling S, Haider U, Rabel K, Coombe N, Jankowicz D, Sockalingam S. Assessing the Impact on Electronic Health Record Burden After Five Years of Physician Engagement in a Canadian Mental Health Organization: Mixed-Methods Study. JMIR Hum Factors. 2025 May 9;12:e65656. doi: 10.2196/65656. PMID: 40344205; PMCID: PMC12083741.
19. Van Veen, D., Van Uden, C., Blankemeier, L. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* **30**, 1134–1142 (2024). https://doi.org/10.1038/s41591-024-02855-5
20. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, Bensahla A, Bjelogrlic M, Lovis C. Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review. J Med Internet Res. 2025 May 15;27:e68998. doi: 10.2196/68998. PMID: 40371947; PMCID: PMC12123242.
21. Croxford E, Gao Y, Pellegrino N, Wong K, Wills G, First E, Liao F, Goswami C, Patterson B, Afshar M. Current and future state of evaluation of large language models for medical summarization tasks. Npj Health Syst. 2025;2:6. doi: 10.1038/s44401-024-00011-2. Epub 2025 Feb 3. PMID: 40124388; PMCID: PMC11928168.
22. Ripp JA, Pietrzak RH, de Guillebon E, Peccoralo LA. Association of clerical burden and EHR frustration with burnout and career intentions among physician faculty in an urban academic health system. Int J Med Inform. 2025 Mar;195:105740. doi: 10.1016/j.ijmedinf.2024.105740. Epub 2024 Dec 1. PMID: 39644795.