Kannan S. Comparative evaluation of large language models in assessing off-label drug use and generating patient consent forms: A Cross-sectional study. Pharmacy Practice 2026 Apr-Jun;24(2):3477.
https://doi.org/10.18549/PharmPract.2026.2.3477

Original Research

Comparative evaluation of large language models in assessing off-label drug use and generating patient consent forms: A Cross-sectional study

Kannan Sridharan

Received (first version): 19-Nov-2025 Accepted: 20-Apr-2026 Published online: 02-Jun-2026

Abstract

Background: Off-label drug use (OLDU) involves the prescription of medications beyond their approved indications, populations, or dosages, and is a common practice in areas such as pediatrics, geriatrics, psychiatry, and oncology. Ethical OLDU necessitates sound clinical justification and transparent, patient-centered consent. With the rise of large language models (LLMs) in healthcare, their utility in supporting OLDU through evidence synthesis and consent documentation remains underexplored. **Methods:** A cross-sectional study was conducted evaluating the performance of two LLMs, ChatGPT-4.0 and DeepSeek-V3, across seven OLDU scenarios encompassing vulnerable populations and emerging indications. Each LLM was prompted to assess OLDU using the BRAvO decision-making framework based on the PrOACT-URL structure and to generate corresponding patient consent forms. Fifteen standardized prompts were used, and responses were evaluated using rubric-based scoring systems for clinical reasoning (maximum score = 27) and consent form quality (maximum score = 33 when reproductive concern was applicable and 30 for others). **Results:** Both LLMs provided structured and complete responses across all scenarios. ChatGPT demonstrated superior readability, empathetic tone, and patient-focused consent forms, consistently scoring in the excellent range. DeepSeek provided higher clinical detail and included extensive guideline references, with strong medico-legal structuring in its documentation. Scenario-specific variations were noted in how each LLM addressed uncertainties, risk mitigation, and reproductive considerations. Overall, both models scored within the excellent range for OLDU assessment and consent generation. **Conclusion:** LLMs such as ChatGPT and DeepSeek can effectively support OLDU decision-making and informed consent processes. Their integration into clinical workflows offers promise, though expert oversight remains essential to ensure accuracy, ethical compliance, and patient-centered care.

Keywords: OLDU, Off-label drug use, Decision-making, Artificial intelligence, AI

INTRODUCTION

Off-label drug use (OLDU) refers to the prescription and administration of medications outside their approved indications, dosages, or patient populations as authorized by regulatory agencies such as the U.S. Food and Drug Administration (FDA)[1]. This practice is pervasive across medical specialties, particularly in neonatology, pediatrics, and oncology, though it extends to all clinical disciplines. While prevalence rates vary, studies estimate that approximately 21% of all prescriptions are for off-label use, reflecting its integral role in modern therapeutics[2]. The implications of OLDU are complex, as it often bridges gaps in treatment where evidence-based alternatives are lacking. A systematic review of randomized clinical trials comparing off-label and approved drug use found that off-label prescriptions were associated with more favorable overall outcomes (odds ratio [OR]: 0.72; 95% confidence interval [CI]: 0.54–0.95) and comparable patient-relevant outcomes (OR: 0.74; 95% CI: 0.56–0.98)[3]. However, the variability in outcomes underscores the need for rigorous evaluation of off-label applications.

Pediatric populations are particularly susceptible to OLDU due to the historical exclusion of children from clinical trials, leading to a reliance on extrapolated data. A large-scale analysis of pediatric prescriptions revealed that 28.1% of patients (779,270 out of 2,773,770) received at least one off-label medication[4]. Recent U.S. estimates indicate an alarming 41.2 million annual off-label prescriptions for children, with a steady increase observed between 2006 and 2015[5]. Similarly, in oncology, off-label use ranges from 13% to 71% in adults and 18% to 41% among inpatients, with 7–31% of these prescriptions lacking support from standard treatment guidelines or drug compendia[6]. While OLDU can be clinically justified, it is not without risks. Serious adverse events, including fatalities, have been documented, highlighting the imperative for careful risk-benefit assessment and patient monitoring[7].

The legal framework governing OLDU remains ambiguous in many jurisdictions. Although laws mandate that such use must meet acceptable standards of care, require full patient disclosure, and necessitate informed consent, they often fail to delineate specific obligations for prescribers[8]. Ethically, OLDU is permissible when it adheres to the core principles of medical ethics: respect for autonomy, beneficence, non-maleficence, and justice[9]. Key criteria for ethical off-label prescribing include evidence-based justification, absence of suitable approved alternatives, a favorable risk-benefit ratio, and alignment with the patient's best interests, all while ensuring the practice is not tantamount to experimental research[9]. The ethical imperative for transparency is further emphasized by the need to maintain

Kannan Sridharan*. Professor, Department of Pharmacology & Therapeutics, College of Medicine & Health Sciences, Arabian Gulf University, Manama, Kingdom of Bahrain. skannandr@gmail.com

www.pharmacypractice.org (eISSN: 1886-3655 ISSN: 1885-642X)

patient trust and uphold shared decision-making.

Clinical pharmacists, in collaboration with clinical pharmacologists and physicians, play a pivotal role in optimizing OLDU. Their expertise in pharmacotherapy positions them as critical stakeholders in institutional decision-making, particularly within hospital pharmacotherapeutics committees and national healthcare advisory bodies[10]. In the U.S., the dynamic between pharmacists and prescribers has been shown to significantly influence indication-based off-label prescribing, underscoring the importance of interdisciplinary communication[11]. Pharmacists contribute to risk mitigation by evaluating drug interactions, monitoring adverse effects, and ensuring adherence to evidence-based protocols. Their involvement is especially crucial in high-risk populations, such as pediatrics and oncology, where off-label use is frequent and the margin for error is narrow.

Large language models (LLMs), a subset of artificial intelligence (AI) based on natural language processing, have emerged as transformative tools in healthcare, capable of assisting with diagnostics, treatment personalization, and patient education[12]. These models, trained on vast datasets, excel at synthesizing complex medical information and generating human-like responses. In prior research, we demonstrated that LLMs can effectively identify prescription errors and contribute to medication review and reconciliation processes, thereby enhancing medication safety[13]. Furthermore, their ability to generate patient-friendly educational materials, encompassing drug indications, therapeutic goals, administration guidelines, and potential side effects, positions them as valuable adjuncts to clinical practice[14].

Given the expanding role of clinical pharmacists in OLDU and the growing adoption of AI in healthcare, there is a pressing need to evaluate the capabilities of LLMs in this domain. Specifically, their potential to assess the rationale for off-label prescriptions and generate comprehensive patient consent forms remains underexplored. Informed consent for OLDU requires clear communication of risks, benefits, and alternatives, a task that demands both accuracy and accessibility. LLMs may streamline this process by drafting tailored consent documents, provided their outputs are rigorously validated for clinical and ethical appropriateness.

This study aims to conduct a comparative evaluation of leading LLMs in two key areas: (1) assessing the evidence-based justification for off-label drug use, and (2) generating patient consent forms that meet regulatory and ethical standards. By examining the performance of models such as ChatGPT-4 and DeepSeek-V3, we seek to elucidate their strengths, limitations, and practical utility in supporting clinicians and pharmacists. The findings will inform the best practices for integrating AI into OLDU workflows while safeguarding patient autonomy and safety.

METHODS

Study design and ethics

A cross-sectional observational study was carried out during May and June 2025. Ethics approval was not required considering the nature of study and we adhered to the latest Declaration of Helsinki guidelines.

Large language models

ChatGPT 4.0: ChatGPT-4.0, developed by OpenAI, is a state-of-the-art large language model based on the GPT-4 architecture. It is a multimodal model capable of processing both text and image inputs and is designed to produce contextually relevant and coherent responses across a wide range of domains, including medical and pharmaceutical contexts. For this study, ChatGPT-4.0 was accessed through the ChatGPT web platform with GPT-4 enabled settings. The model has been extensively trained on publicly available datasets and licensed sources, enabling it to simulate human-like dialogue and provide detailed, structured information, including clinical decision support, ethical reasoning, and documentation relevant to off-label drug use. While the model lacks real-time internet access during response generation, it leverages its vast training corpus and reinforcement learning-based tuning to generate responses that align with current medical knowledge and practice standards up to its knowledge cut-off in April 2023.

DeepSeek-V3: DeepSeek-V3 is an advanced LLM developed by DeepSeek, designed to process and generate human-like text responses across a wide range of domains, including healthcare and clinical research. In this study, DeepSeek-V3 was evaluated alongside other LLMs for its ability to assess the rationale for off-label drug use and generate patient consent forms. The model was accessed via its publicly available interface during the study period (January–March 2024) and was prompted using standardized inputs to ensure consistency in response generation. DeepSeek-V3's performance was assessed based on its capacity to provide evidence-based justifications for off-label prescriptions, synthesize complex medical information, and produce clear, ethically compliant consent documentation. As a state-of-the-art LLM, it leverages a transformer-based architecture trained on extensive biomedical and pharmacological literature, enabling it to interpret clinical contexts and generate contextually appropriate outputs. Its responses were evaluated for accuracy, completeness, readability, and adherence to ethical guidelines governing off-label drug use.

OLDU scenarios

Seven OLDU case scenarios spreading across different domains such as age groups (pediatric, adults, and older adults), vulnerable population (pregnant, lactating women, and mental illness), and status of OLDU (established and emerging) themes as outlined in Table 1 were assessed in this study.

Framework for assessing OLDU

The Benefit and Risk Assessment for Off-label use (BRAvO) decision-making tool[15], which utilizes the PrOACT-URL framework[16], to systematically evaluate off-label drug applications was used in this study. While this practical framework was initially designed for pediatric populations to provide a structured evaluation of therapeutic benefits

https://doi.org/10.18549/PharmPract.2026.2.3477

Table 1. Details about the OLDU scenarios used in this study

Scenario title	OLDU domain	Scenario
Metoclopramide for Hyperemesis Gravidarum	Pregnant	A 26-year-old primigravida at 10 weeks gestation presents with severe hyperemesis gravidarum, manifesting as intractable nausea and vomiting that has led to significant weight loss of 4.3 kg, moderate dehydration, and emerging electrolyte disturbances. The patient has failed initial conservative management strategies, including dietary modifications, vitamin B6 supplementation, and initial antiemetic interventions. The clinical team is considering the off-label use of metoclopramide (10 mg three times daily orally) as a therapeutic intervention.
Furosemide for Chronic Lung Disease in Preterm Neonates	Neonates	A 28-week gestation male neonate, weighing 1,100 grams, is admitted to the neonatal intensive care unit following a complicated preterm delivery with respiratory distress syndrome and evolving bronchopulmonary dysplasia. Despite standard respiratory management including surfactant administration, mechanical ventilation, and corticosteroid therapy, the infant demonstrates persistent pulmonary edema, increased work of breathing, and radiographic evidence of progressive chronic lung disease. The clinical team is considering off-label furosemide administration.
Sertraline for Post-Partum Depression in a Lactating Mother	Lactating mother	A 32-year-old multiparous woman, who delivered a healthy full-term infant four weeks ago, presents with severe postpartum depression characterized by persistent depressive symptoms, significant anxiety, social withdrawal, disrupted maternal-infant bonding, and emerging suicidal ideation without active plan. Her comprehensive psychiatric evaluation reveals a complex clinical picture with moderate-to-severe depressive symptoms that have not sufficiently responded to initial non-pharmacological interventions including psychotherapy, support group engagement, and lifestyle modifications. The clinical team proposes off-label sertraline administration.
Gabapentin for Generalized Anxiety Disorder	Psychiatric condition	A 45-year-old female presents with generalized anxiety disorder characterized by persistent and debilitating worry, significant autonomic hyperarousal, sleep disturbances, and substantial functional impairment across occupational and social domains. Despite initial standard treatment approaches including cognitive behavioral therapy and selective serotonin reuptake inhibitors, the patient demonstrates inadequate symptom control and limited therapeutic response. After comprehensive psychiatric and neurological consultation, the clinical team proposes off-label gabapentin (300 mg orally daily).
Spironolactone for Cirrhotic Edema	Established OLDU	A 58-year-old male with end-stage alcoholic liver cirrhosis presents with refractory ascites, demonstrating Child-Pugh class C classification and significant fluid accumulation that has become increasingly resistant to standard diuretic management with furosemide and standard-dose spironolactone. Despite optimal medical management, the patient experiences recurrent large-volume paracentesis requirements, progressive nutritional compromise, and increasing intra-abdominal pressure causing significant functional impairment. The hepatological team considers an aggressive off-label approach utilizing high-dose spironolactone at 500 mg daily.
Estradiol Patch for GAHT	Emerging OLDU	A 28-year-old transgender woman presenting for gender-affirming hormone therapy (GAHT) seeks comprehensive medical management to align her physiological characteristics with her gender identity. Following extensive psychological evaluation, informed consent discussions, and baseline endocrinological assessments, the clinical team proposes initiating estradiol transdermal patch (0.1 mg/day) as off-label intervention as a part of her hormonal transition strategy.
Metformin as Anti-aging drug	Emerging OLDU and older adults	A 72-year-old apparently healthy male with no diabetes presents for comprehensive geriatric wellness evaluation, demonstrating increasing interest in emerging anti-aging interventions. The patient exhibits early signs of age-associated metabolic dysfunction, including mild insulin resistance, subtle inflammation markers, and preliminary molecular aging indicators identified through advanced comprehensive metabolic and genomic screening. After extensive discussion regarding potential off-label pharmacological interventions targeting fundamental aging mechanisms, the clinical team proposes metformin administration (500 mg daily) as a potential anti-aging strategy.

OLDU: Off-label drug use; and GAHT: Gender-affirming hormone therapy.

versus potential risks, incorporating clinical pharmacological principles for age-appropriate dosing, its utility extends to other vulnerable patient groups. Specifically, it can inform off-label prescribing decisions in special populations such as pregnant women, elderly patients, individuals with obesity, and critically ill patients, where off-label drug use is particularly prevalent. By offering a standardized approach to benefit-risk assessment, this framework supports clinicians in optimizing both the efficacy and safety of pharmacotherapy in scenarios where approved treatment options may be limited or unavailable. The domains under which OLDU was assessed with BRAvO using PrOACT-URL framework (Supplementary File 1) are as follows:

- Problem and alternatives
- Objectives: Efficacy
- Objectives: Safety
- Objectives: Dosing
- Consequences
- Tradeoffs
- Uncertainties
- Risk tolerance and
- Linked Decisions

LLM prompts

A total of 15 prompts were used in this study for each LLM (Supplementary File 2). Briefly, they are as follows:

- Prompt 1: Related to essential elements to be present in the informed consent document related to OLDU. Consent documents from three hospitals related to OLDU[17-19] were uploaded in this prompt.
- Prompts 2, 4, 6, 8, 10, 12, and 14 were related to the OLDU scenarios.

https://doi.org/10.18549/PharmPract.2026.2.3477

- Prompts 3, 5, 7, 9, 11, 13 and 15 were related to generating informed consents related to the respective OLDU scenarios.
- The veracity of the outputs was assessed individually according to the rubrics that are scored between 0 and 3 for each element in the BRAvO using PrOACT-URL framework (Supplementary File 3). Interpretation of the final scores of the OLDU assessment was carried out as follows:
- Excellent (24–27 points): Output is comprehensive, well-structured, and demonstrates a high level of clinical reasoning across all domains. It thoroughly addresses each element of the BRAvO-PrOACT-URL framework with contextualized insights, appropriate medical terminology, and aligned with clinical guidelines. Suitable for high-stake educational, clinical, or publication use.
- Good (18–23 points): Output is mostly complete and understandable, with moderate clinical depth. Some domains may be briefly addressed or moderately justified, but the overall structure is sound. It demonstrates a clear grasp of the framework with minor areas requiring clarification or elaboration.
- Fair (12–17 points): Output shows basic understanding but lacks sufficient depth in several key areas such as safety, dosing, or ethical decision-making. Content may be superficial, general, or not fully aligned with clinical standards. Requires expert revision before practical application.
- Poor (<12 points): Output is incomplete or inaccurate, with major gaps in reasoning and failure to adequately address the BRAvO-PrOACT-URL domains. Key elements such as efficacy rationale or safety concerns may be missing or misrepresented. Not suitable for clinical or educational use without major corrections.

Similarly, the generated consent forms were assessed using the rubrics for that are scored between 0 and 3 for each essential element (Supplementary File 4). Interpretation of the final scores of the consent document was carried out as follows:

- Excellent (27–33 points): Consent document is comprehensive, ethically robust, and written in a highly patient-friendly and context-sensitive manner.
- Good (20–26 points): Consent is mostly complete and understandable with only minor issues in personalization or tone.
- - Fair (13–19 points): Several key elements are vague or missing; readability or risk-benefit explanation may be insufficient.
- Poor (<13 points): Consent lacks essential ethical or informational components and may not support valid informed consent.

In case the reproductive considerations were not applicable, each category was reduced by three points for the assessment of patient consent form. The scores were independently assessed by two investigators and any discrepancies were resolved through discussion.

RESULTS

Both the LLMs provided responses to all scenarios and generated patient consent documents. Detailed analysis of their response based on each scenario is as follows:

Scenario 1 (Metoclopramide for hyperemesis gravidarum)

Both the LLMs provided structured response to the domains listed under the framework for assessing OLDU (Supplementary Files 5 and 6; responses to prompt number 2) but variations were observed in terms of their clarity, depth, and patient-centeredness. In general, ChatGPT provided an excellent response (score=25) compared to DeepSeek that provided a fair response (score=17) (Table 2). The responses from ChatGPT were more coherent and inclusive to each domain while the DeepSeek listed the appropriate references at the end and has considered the possibility of theoretical adverse event for metoclopramide in neonates. ChatGPT provided a well-organized, criterion-wise analysis under the BRAvO

Table 2. Evaluation of LLMs on the OLDU assessment using BRAvO framework

Domains of ICF	Metoclopramide in hyperemesis gravidarum		Furosemide in preterm neonates		Sertraline in lactating mother		Gabapentin for generalized anxiety disorder		High dose spironolactone for cirrhotic edema		Estradiol patch for GAHT		Metformin as anti-aging	
	C	D	C	D	C	D	C	D	C	D	C	D	C	D
Problem and alternatives	3	2	3	3	3	3	3	2	3	3	3	3	3	3
Objectives: Efficacy	3	2	3	2	3	2	3	2	3	3	3	3	3	2
Objectives: Safety	3	2	3	2	2	2	2	3	3	3	2	3	3	3
Objectives: Dosing	3	2	3	3	3	3	3	3	3	3	3	3	3	3
Consequences	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Tradeoffs	3	3	3	3	3	2	3	3	3	3	3	3	3	3
Uncertainties	2	2	3	3	3	3	3	3	3	3	3	3	3	3
Risk tolerance	2	2	3	3	3	3	3	3	3	3	3	3	2	3
Linked decisions	3	1	3	1	3	3	3	3	3	3	3	3	3	3

C: ChatGPT; D: DeepSeek; and GAHT: Gender-affirming hormonal therapy.

https://doi.org/10.18549/PharmPract.2026.2.3477

framework, closely aligning with clinical decision-making processes, comprehensively addressing all the elements such as unmet need, drug licensing status, alternatives, and safety profiles with specific references to pregnancy safety data, clinical guidelines, and real-world usage. Additionally, ChatGPT identified the clinician's judgment in weighing tradeoffs, uncertainties, and linked decisions. The benefit-risk assessment was clearly articulated and acknowledged both clinical and ethical considerations. However, the potential fetal risks and any adverse events (when drugs are used for off-label purpose) were not mentioned as one of the uncertainties by ChatGPT. On the other hand, DeepSeek carried out a systematic structural evaluation that includes guideline-based reasoning, comparative analysis of therapeutic alternatives, risk assessment, and patient context. Its clinical reasoning was sound and incorporates key reference citations (such as ACOG Bulletin, Einarson meta-analysis), indicating a strong knowledge base which was not observed with ChatGPT. However, some framework elements such as analog use, efficacy, dosing simulations and real-time monitoring strategies were less explicitly addressed compared to ChatGPT as well as did not mentioning the lack of randomized clinical trial evidence to be considered under the uncertainty's domain, and the way for disseminating the observations following OLDU.

The LLMs generated consent forms (Supplementary Files 5 and 6; responses to prompt number 3) are appropriate, structured and incorporated the key elements. The ChatGPT's response was good (score=25) was more readable and patient-friendly with an empathetic tone, making it very appropriate for clinical use and enhancing informed decision-making. Additionally, it included placeholders for institutional details and offers explicit declarations of understanding and consent. Similarly, the consent form generated by DeepSeek was good (score=21) and was comprehensive and legally detailed, with clear sections on diagnosis, drug rationale, risks, alternatives, and a structured signature section. It included physician attestation and reference citations and that a copy of this consent form has been handed over to the patient making it robust for medico-legal documentation. However, the language was more formal and medicalized, potentially requiring simplification for low-literacy populations. Further, DeepSeek did not include monitoring plans, never provided opportunity to ask questions or refuse to accept the treatment.

Scenario 2 (Furosemide for chronic lung disease in preterm neonates)

Both the LLMs provided structured, and methodically addressing the core expected elements for this OLDU (Supplementary Files 5 and 6; responses to prompt 4). ChatGPT's response was excellent (score=27; Table 2), clinically grounded, emphasizing real-world contextual judgment, such as risk mitigation strategies, precedents in tertiary care, acknowledging uncertainties (especially long-term effects), and incorporates team-based, multidisciplinary decision-making as part of its risk management. Similarly, DeepSeek's response was excellent (score=23; Table 2), technically more detailed, referenced appropriate neonatal consensus guidelines, stated explicit goals such as oxygen index improvement and ventilator weaning, provided potential comparison of outcomes with other diuretic options, with the slightly more formal and academic in tone, suggesting its utility in documentation and clinical reporting rather than direct communication with lay audiences. However, DeepSeek did not provide explicit information on the analog use, the relevance of mechanism of action to the concerned OLDU, and the way for dissemination of findings.

Regarding the informed consent documents, both the LLM included most of the essential elements (Supplementary Files 5 and 6; responses to prompt 5). ChatGPT's response was excellent (score=26; Table 3) but did not mention the

Table 3. Summary of evaluation of LLMs on the generated consent forms

Domains of ICF	Metoclopramide in hyperemesis gravidarum		Furosemide in preterm neonates		Sertraline in a lactating mother		Gabapentin for generalized anxiety disorder		High dose spironolactone for cirrhotic edema		Estradiol patch for GAHT		Metformin as anti-aging	
	C	D	C	D	C	D	C	D	C	D	C	D	C	D
Purpose and context	3	2	3	2	3	3	3	3	3	3	3	3	3	3
Drug description	3	3	3	2	3	3	3	3	3	3	2	3	3	3
Off-label explanation	3	2	3	2	3	2	3	2	2	2	2	1	3	2
Benefits	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Risks and side effects	2	3	3	3	3	3	2	3	3	3	3	3	3	3
Alternatives	2	3	3	3	3	3	3	3	3	3	2	3	3	3
Monitoring plan	1	0	0	3	1	3	0	3	2	3	2	3	3	3
Reproductive consideration (if applicable)	2	2	NA											
Patient autonomy	2	1	3	3	2	2	3	2	3	3	3	3	3	3
Readability and tone	3	0	2	0	3	0	3	1	3	3	3	1	3	3
Documentation elements	3	3	3	3	3	3	3	3	3	3	3	3	3	3

C: ChatGPT; D: DeepSeek; ICF: Informed consent form; GAHT: Gender-affirming hormonal therapy; and NA: Not applicable.

monitoring plan and occasionally included technical jargon. Similarly, DeepSeek's response was good (score=22; Table 3) but did not describe adequately the OLDU as well as furosemide. ChatGPT's tone was empathetic and suitable for neonatal intensive care unit caregivers, potentially enhancing parent understanding and aligning with principles of shared decision-making and empowerment. DeepSeek on the other hand detailed the daily monitoring plans, provided appropriate reference and stated that a copy of this consent form has been handed over to the patients, but was more technical and used overly technical medical jargons.

Scenario 3 (Sertraline in lactating mother)

Both the LLM's were able to provide a structured response to all the domains in the framework for this scenario (Supplementary Files 5 and 6; responses to prompt 6). Their responses were excellent (ChatGPT score=26; and DeepSeek score=24; Table 2). However, both LLMs did not explicitly state one of the key risk mitigation strategies related to drug ingestion during the time when newborn is expected to sleep, thus reducing the risk of milk transfer. DeepSeek additionally did not state the relevance of mechanism of action of the drug to the OLDU and its effect on disease analog. However, DeepSeek provided an exhaustive list of potential alternatives (cognitive behavioral therapy, electroconvulsive therapy, repetitive transcranial magnetic stimulation and interpersonal psychotherapy), contingency plan in case of failure of OLDU, cited pharmacokinetic details that included milk-to-plasma ratio, and provided more appropriate and relevant references than ChatGPT.

Regarding the patient consent forms, both LLMs generated documents (Supplementary Files 5 and 6; responses to prompt 7). Their responses were excellent (ChatGPT score=27; and DeepSeek score=25; Table 3). However, ChatGPT's description on the monitoring plan, the right to refuse and ask questions were not explicitly stated. DeepSeek on the other hand has not adequately explained the purpose of OLDU and the right to refusal, and the document was overly technical including several medical jargons. But DeepSeek has provided detailed monitoring plans, checklists for symptoms, dosage, side effects, and monitoring, and more relevant references compared to ChatGPT. On the other hand, ChatGPT has maintained an empathetic tone, emphasizing shared decision-making, making it more suitable for real-world patient encounters.

Scenario 4 (Gabapentin for generalized anxiety disorder)

The responses of LLMs on this scenario can be observed in Supplementary Files 5 and 6 (responses to prompt 8). Their responses were excellent (ChatGPT score=26 and DeepSeek score=25; Table 2). ChatGPT's response was pragmatic, outlining the unmet medical need, acknowledging gabapentin's lack of FDA approval for this indication, incorporating pharmacologic rationale, safety profile, dosing guidance, and patient-specific considerations such as absence of substance use history. However, it did not specify suicidal ideations as one of the safety concerns. DeepSeek's evaluation on the other hand was more technical and included both pharmacological and guideline-based references, explicitly comparing gabapentin's tradeoffs against alternatives, mentioned suicidal ideations (though erroneously as black box warning), tapering protocol, and contingency plan in case of gabapentin failure.

The patient consent documents generated by the LLMs are outlined in the Supplementary Files 5 and 6 (response to prompt 9). ChatGPT's consent form was excellent (score=26; Table 3), easily comprehensible with an empathetic tone but did not mention the risk of suicidal ideations. DeepSeek's response was also rated excellent (score=26; Table 3) but was more formal, containing technical details, included checklists, provided a comprehensive breakdown of side effects (common, less common, serious), and explicit instructions on usage and follow-up. DeepSeek also mentioned suicidal ideations and quite rightly included 24-hour emergency contact details.

Scenario 5 (High-dose spironolactone for cirrhotic edema)

The LLM responses to this scenario can be found in Supplementary Files 5 and 6 (responses to prompt 10). Both the LLMs were excellent (score=27; Table 2) and were structured, coherent, and in alignment with the framework for OLDU. Both LLMs emphasized the balance between potential benefits, such as symptom relief and reduction in invasive procedures, and known risks, including hyperkalemia and renal dysfunction, with appropriate strategies for risk mitigation. ChatGPT highlighted the mechanistic and clinical reasoning for using high-dose spironolactone in highly refractory cases. DeepSeek on the other hand discussed more alternatives, risk mitigation and monitoring strategies than ChatGPT. Additionally, DeepSeek has also mentioned contingency plans in case of extreme hyperkalemia.

The patient consent forms generated from the LLMs can be observed in Supplementary Files 5 and 6 (responses to prompt 11). Both their responses were excellent (ChatGPT score=25; and DeepSeek score=26; Table 3). The consent form provided by ChatGPT is written at an accessible reading level, including a plain-language explanation of the rationale for high-dose use with an empathetic tone. DeepSeek has provided extensive monitoring plan, and adverse effects in the form of checklists, was formal and exhaustive, and provided emergency contact details. However, the form had more legal tone making it ideal for medico-legal compliance and hospital-based implementation. However, the language and structure may be less accessible for patients with low health literacy without additional explanation from a clinician.

Scenario 6 (Estradiol patch for GAHT)

Supplementary Files 5 and 6 (responses to prompt 12) contain responses of LLM for this scenario. Their responses were excellent (ChatGPT score=26; and DeepSeek score=27; Table 2). ChatGPT did not provide risk mitigation strategies while DeepSeek identified gender dysphoria as a separate disease entity and provided International Classification of Disease classification code (ICD-10 F64.0), suggests adding anti-androgens, provides appropriate monitoring strategies and contingency plans in case of treatment failure to estradiol gel and add on anti-androgens.

The generated patient consent forms can be found in Supplementary Files 5 and 6 (responses to prompt 13). Their responses were excellent (ChatGPT score=29; and DeepSeek score=26; Table 3). ChatGPT failed to describe OLDU and estrogen patch completely and did not provide adequate monitoring plans. However, ChatGPT's consent form reflected a high degree of inclusivity and readability and included all core elements in an affirming language that supports gender identity and autonomy, making it more suitable in a clinical or community-based setting. Similarly, DeepSeek did not describe about OLDU adequately and had more technical jargons but mentioned the appropriate timelines for each feminization effect, explicitly details checklists for adverse events and monitoring parameters, fertility and reproductive health including infertility and sperm preservation, and contact details of local support group and trans lifeline. DeepSeek emphasized clinical rigor and included highly comprehensive details that may be more appropriate for specialist institutional documentation.

Scenario 7 (Metformin as anti-aging)

The responses of the LLMs to this OLDU scenario can be observed in Supplementary Files 5 and 6 (responses to prompt 14). Both the LLMs responded excellently (score=26; Table 2). ChatGPT's response demonstrated a clear, logical structure across all domains of the framework with a balanced discussion of unmet needs, including markers of metabolic aging, and provided the mechanistic rationale for using metformin based on its activation of adenosine monophosphate-activated protein kinase and inhibition of mechanistic target of rapamycin. DeepSeek provided a more technical and evidence-oriented narrative, integrating the role of specific biomarkers such as homeostatic model assessment of insulin resistance and highly sensitive C-reactive protein, provided wider alternatives such as senolytics and nicotinamide adenine dinucleotide precursors, emphasized that metformin for this OLDU was not a standard of care and realistic goal setting, distinguishing between theoretical aging delay and actual life extension, thus underscoring ethical clarity.

The patient consent forms generated by LLMs are present in Supplementary Files 5 and 6 (response to prompt 15). The consent forms were excellent (ChatGPT score=30; and DeepSeek score=29; Table 3). ChatGPT's consent form was explicitly written explaining the off-label nature of the intervention, its proposed benefits, known risks, alternatives, and patient rights in a tone that is both empathetic and empowering. DeepSeek on the other hand provided an exhaustive and formatted consent form detailing treatment purpose, alternatives, known risks, investigational status, exit strategies, and follow-up plans. DeepSeek stated providing a patient copy of the consent form.

DISCUSSION

Key findings

The comparative evaluation of LLMs in this study revealed that both ChatGPT and DeepSeek can generate clinically relevant assessments and ethically sound informed consent forms for OLDU scenarios across diverse populations and therapeutic contexts. ChatGPT consistently demonstrated a strong orientation toward patient-centered communication, with high readability, empathetic tone, and clear structuring of information, making it particularly well-suited for direct patient interactions and shared decision-making. In contrast, DeepSeek offered technically rigorous responses with greater incorporation of guideline-based evidence, pharmacological references, and structured monitoring strategies, reflecting a more formal and documentation-oriented approach. While both LLMs performed excellently across most scenarios in applying the BRAvO using PrOACT-URL framework, ChatGPT slightly outperformed DeepSeek in readability and contextual inclusivity, particularly in consent form generation. DeepSeek, however, provided superior depth in medical detailing, contingency planning, and institutional-level formatting. These findings underscore the complementary strengths of each LLM, with ChatGPT excelling in patient engagement and DeepSeek in clinical thoroughness and medico-legal robustness, thereby supporting their potential integration into OLDU workflows tailored to specific healthcare settings and user needs.

Comparison with existing literature

OLDU remains a prevalent and often indispensable component of clinical practice, particularly in scenarios where therapeutic options are limited or emerging evidence supports alternative uses of approved medications. In an era marked by rapidly evolving treatment paradigms and the frequent absence of universally accepted clinical guidelines, making informed decisions regarding OLDU is inherently complex and necessitates a systematic, multidisciplinary approach[20]. A recent scoping review of 31 published guidance documents on OLDU emphasized several best practices to navigate this complexity. These included the importance of grounding decisions in robust scientific evidence, assembling diverse expertise for evaluating and synthesizing that evidence, employing structured processes to guide therapeutic recommendations, integrating OLDU with timely and meaningful clinical research (including real-world data), and fostering collaborative networks among clinicians, researchers, regulatory bodies, policymakers, and sponsors to ensure both effective implementation and iterative evaluation of OLDU strategies[21].

Accountable and ethical prescribing of off-label medications must be underpinned by strict adherence to evidence-based practice, thorough assessment of patient-specific factors, and a judicious evaluation of the expected benefit-risk ratio. A central element in this process is the delivery of a well-informed, ethically robust consent document that clearly communicates the rationale, uncertainties, and available alternatives to the patient [22]. The current study demonstrates that LLMs, especially ChatGPT 4.0, possess the capability to construct such clinical rationale with mechanistic insights and contextual clarity. Notably, ChatGPT 4.0 often framed OLDU as a logical extension of a drug's approved use, highlighting therapeutic proximity and analogous pathophysiological considerations, an approach consistent with real-world OLDU rationale and regulatory perspectives.

However, concerns about inappropriate or unregulated OLDU remain, particularly in cases where such use may divert essential medications from populations with on-label indications, leading to shortages and affecting standard care delivery[23]. In this regard, the present study's findings are reassuring. For example, DeepSeek explicitly recognized the use of metformin for anti-aging as a non-standard intervention, raising important cautionary flags regarding inappropriate expansion of OLDU without sufficient evidence base or regulatory support. This reflects the model's ability to identify areas of ethical ambiguity and align responses with current clinical norms. Both LLMs, and especially DeepSeek, demonstrated an impressive capacity to enumerate and analyze therapeutic alternatives across the studied scenarios, including a comparative evaluation of their advantages and disadvantages. This capability is invaluable in clinical decision-making, as it supports a more nuanced and individualized approach to OLDU by aiding prescribers in contextualizing the appropriateness of off-label choices.

Within institutional frameworks, clinical pharmacists play an instrumental role in shaping and governing OLDU policies through their contributions to Drug and Therapeutics Committees (DTCs). These professionals are tasked with safeguarding patient welfare by ensuring that therapeutic decisions are grounded in scientific rigor and economic responsibility, thus protecting healthcare systems from adopting costly treatments with marginal benefit[24]. The promising results from this study suggest that LLMs can be effectively integrated into these committee processes. Building on previous research that demonstrated the ability of LLMs to identify ethical issues in clinical research protocols[25], this study extends their utility to the domain of OLDU evaluation. LLMs may serve as educational and decision-support tools for newly appointed DTC members by simulating expert-level reasoning and reinforcing standard frameworks such as BRAvO-PrOACT-URL. Furthermore, DTCs are increasingly responsible for conducting prescription-indication audits as part of routine pharmacoepidemiologic surveillance to detect and characterize OLDU patterns within hospital settings[26]. The prompts and assessment framework utilized in this study can serve as a blueprint for DTCs to evaluate the appropriateness of specific OLDU practices systematically. Additionally, the study highlights the potential utility of LLMs in generating informed consent documents. Both ChatGPT and DeepSeek produced consent forms that generally met ethical and informational standards, although the tone and complexity differed. DeepSeek's outputs leaned toward formal, medico-legal precision, whereas ChatGPT offered more empathetic, patient-accessible language. In high-volume hospital environments where clinicians face significant time constraints, DTCs may be assigned the task of preparing standardized consent templates for OLDU, a task in which LLMs could offer efficient, scalable, and customizable solutions.

Overall, this study supports the integration of LLMs as adjuncts to clinical judgment in the ethical, evidence-informed, and patient-centered deployment of off-label therapies. Their structured approach to decision-making, versatility in communication, and capacity to emulate professional standards suggest a future in which LLMs can meaningfully contribute to OLDU governance, documentation, and patient engagement. However, continued oversight by expert clinicians, pharmacists, and ethicists remains crucial to ensure that the use of these technologies' complements, rather than replaces, human expertise.

Strengths, limitations and future directions

The key strength of this study lies in its comprehensive, structured, and rubric-based evaluation of two state-of-the-art large language models across a diverse array of OLDU scenarios. By employing the BRAvO using PrOACT-URL framework and standardized scoring rubrics for informed consent, the study ensures an objective, reproducible, and clinically grounded comparison that reflects real-world considerations. The inclusion of seven distinct OLDU cases spanning pediatric, adult, geriatric, pregnant, lactating, and transgender populations allows for a nuanced analysis of the LLMs' capabilities across ethically sensitive and medically complex contexts. Additionally, the dual focus on evidence-based assessment and patient consent generation mirrors the twin responsibilities of clinicians, scientific justification and ethical communication, thereby enhancing the translational value of the findings for clinical practice.

Despite these strengths, the study has several limitations. The evaluation of LLM outputs was based on pre-defined prompts, and while this standardized the comparison, it may not fully capture the models' dynamic adaptability in unstructured, real-time clinical interactions. Furthermore, the LLMs assessed do not have access to real-time medical databases or evolving clinical guidelines beyond their training cutoffs, potentially limiting the currency of their recommendations. The scoring system, although rigorous, involved subjective judgment, which may introduce reviewer bias despite consensus resolution methods. Moreover, the performance of LLMs in generating consent forms was evaluated for content accuracy and readability but not validated in actual patient encounters, leaving questions about real-world comprehension and acceptance by diverse patient populations. However, despite these limitations, for clinicians, these findings suggest that LLMs can serve as valuable adjuncts in OLDU decision-making and documentation, particularly in streamlining initial drafts of benefit-risk assessments and patient consent forms. However, the output must be critically reviewed and tailored to individual patient contexts to ensure clinical appropriateness and legal defensibility. Future research should focus on expanding the scope of evaluation to include more LLMs, real-time testing in clinical settings, and integration with electronic health records to assess workflow efficiency and patient outcomes. Additionally, incorporating patient and caregiver feedback on the acceptability and clarity of LLM-generated consent forms will be crucial in refining these tools for frontline use. Developing fine-tuned, domain-specific LLMs with regular updates from curated clinical databases and integrating interdisciplinary oversight could further enhance the safety, accuracy, and ethical robustness of AI-assisted OLDU practices.

CONCLUSION

In conclusion, this study demonstrates that LLMs such as

ChatGPT and DeepSeek hold considerable promise in potentially supporting clinicians with the assessment and documentation of OLDU across diverse clinical scenarios. Both models effectively applied a structured decision-making framework to evaluate therapeutic justification and generated informed consent forms that met key ethical and regulatory standards. While ChatGPT excelled in patient-centered communication and accessibility, DeepSeek provided superior clinical detail and medico-legal formality, highlighting their complementary strengths. These findings underscore the potential of LLMs as supportive tools in promoting evidence-based, ethically sound, and patient-informed OLDU practices. However, their output should be viewed as preliminary drafts requiring expert clinical oversight. As AI technologies continue to evolve, their integration into OLDU workflows must be accompanied by ongoing validation, interdisciplinary collaboration, and robust ethical governance to ensure they enhance, rather than replace, clinician judgment and patient engagement.

CONFLICTS OF INTERESTS

The author have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

References

1. Gota V, Divatia JV. Off-label use of drugs: An evil or a necessity? Indian J Anaesth. 2015 Dec;59(12):767-8.
2. Radley DC, Finkelstein SN, Stafford RS. Off-label prescribing among office-based physicians. Arch Intern Med. 2006;166(9):1021–1026.
3. Ladanie A, Ioannidis JPA, Stafford RS, Ewald H, Bucher HC, Hemkens LG. Off-label treatments were not consistently better or worse than approved drug treatments in randomized trials. J Clin Epidemiol. 2018 Feb;94:35-45.
4. Yackey K, Stukus K, Cohen D, Kline D, Zhao S, Stanley R. Off-label Medication Prescribing Patterns in Pediatrics: An Update. Hosp Pediatr. 2019 Mar;9(3):186-193.
5. Hoon D, Taylor MT, Kapadia P, Gerhard T, Strom BL, Horton DB. Trends in off-label drug use in ambulatory settings: 2006–2015. Pediatrics 2019; 144:e20190896.
6. Saiyed MM, Ong PS, Chew L. Off-label drug use in oncology: a systematic review of literature. J Clin Pharm Ther. 2017 Jun;42(3):251-258.
7. Sadjadi R, Cogdell E, Mostafa ME, Anatelli F, Ackerman L, Wijarnpreecha K, Han MAT. Drug Reaction With Eosinophilia and Systemic Symptoms and Severe Drug-Induced Liver Injury After Off-Label Zonisamide Use for Weight Loss. ACG Case Rep J. 2025 May 24;12(6):e01715.
8. Cook RJ. Off-label drug use as a consent and health regulation issue in New Zealand. J Bioeth Inq. 2015 Jun;12(2):251-8.
9. Zheng Z, Yang M, Wu J. Ethical Off-label Drug use: Need for a Rethink? Indian Pediatr. 2017 Jun 15;54(6):447-450.
10. Schrier L, Hadjipanayis A, Stiris T, Ross-Russell RI, Valiulis A, Turner MA, Zhao W, De Cock P, de Wildt SN, Allegaert K, van den Anker J. Off-label use of medicines in neonates, infants, children, and adolescents: a joint policy statement by the European Academy of Paediatrics and the European society for Developmental Perinatal and Pediatric Pharmacology. Eur J Pediatr. 2020 May;179(5):839-847.
11. Basak R, Bentley JP, McCaffrey DJ 3rd, Bouldin AS, Banahan BF 3rd. The role of perceived impact on relationship quality in pharmacists' willingness to influence indication-based off-label prescribing decisions. Soc Sci Med. 2015 May;132:181-9.
12. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, Zhang H, Liu D, Wang J, Wang X, Gao J, Wang YG, Shao C, Wang W, Li J, Zheng MQ, Yang Y, Tang YD. Revolutionizing Health Care: The Transformative Impact of Large Language Models in Medicine. J Med Internet Res. 2025 Jan 7;27:e59069.
13. Sridharan K, Sivaramakrishnan G. Unlocking the potential of advanced large language models in medication review and reconciliation: A proof-of-concept investigation. Explor Res Clin Soc Pharm. 2024 Aug 17;15:100492.
14. Sridharan K, Sivaramakrishnan G. Investigating the capabilities of advanced large language models in generating patient instructions and patient educational material. Eur J Hosp Pharm. 2024 Dec 30:ejhpharm-2024-004245.
15. van der Zanden TM, Mooij MG, Vet NJ, Neubert A, Rascher W, Lagler FB, Male C, Grytli H, Halvorsen T, de Hoog M, de Wildt SN. Benefit-Risk Assessment of Off-Label Drug Use in Children: The Bravo Framework. Clin Pharmacol Ther. 2021 Oct;110(4):952-965.
16. PROTECT. Pharmacoepidemiological research on outcomes of therapeutics by a European consortium.
17. Checklist for written consent: Unregistered use of medicine.
18. Off-label drugs. Educational material and consent form.
19. Off label use of drug or device. (accessed on May 27, 2025).
20. Lehman K, Aroney E. A guided framework for assessing off-label medication use in psychiatry. Australasian Psychiatry. 2024 Feb;32(1):63-7.
21. Gazarian M, Horton DB, Carleton B, Kinlaw AC, Bushnell GA, Czaja AS, Durrieu G, Gorman EF, Titievsky L, Zito J, Slaughter JL, dosReis S. Optimizing therapeutic decision-making for off-label medicines use: A scoping review and consensus

www.pharmacypractice.org (eISSN: 1886-3655 ISSN: 1885-642X)

recommendations for improving practice and research. Pharmacoepidemiol Drug Saf. 2023 Nov;32(11):1200-1222.
22. Morden NE, Schwartz LM, Fisher ES, et al. Accountable prescribing. N Engl J Med 2013; 369(4): 299–302.
23. Day RO. Ongoing challenges of off-label prescribing. Aust Prescr 2023;46:86-9.
24. Sofat R, Cremers S, Ferner RE. Drug and therapeutics committees as guardians of safe and rational medicines use. Br J Clin Pharmacol. 2020 Jan;86(1):10-12.
25. Sridharan K, Sivaramakrishnan G. Leveraging artificial intelligence to detect ethical concerns in medical research: a case study. Journal of Medical Ethics. 2025 Feb 1;51(2):126-34.
26. Vargas-Rivas JE, Sabater-Hernández D, Calleja-Hernández MA, Faus MJ, Martínez-Martínez F. Role of the hospital pharmacy and therapeutics committee in detecting and regulating off-label drug use. International Journal of Clinical Pharmacy. 2011 Oct;33:719-21.