Title Page

Navigating Diagnostic Challenges in Lymphoma: Patient Narrative Analysis Using Large Language Models

Running title

Patient-reported diagnostic challenges in lymphoma

Author(s)

Feng He

Affiliation

Independent Researcher, Singapore

Correspondence

Feng He

Independent Researcher

Email: feng.he@u.nus.edu

Word count: 3037

Figure/Tables: 5

Version: September 9th, 2025

ABSTRACT

Background: Accurate diagnosis is a recognized challenge in lymphoma, yet little is known about how patients navigate their treatment journey when experiencing diagnostic error.

Methods: We retrospectively analysed patient narratives from house086.com, the largest Chinese online lymphoma forum (2011-2025). Posts related to potential diagnostic errors were retrieved and processed using an AI-based pipeline integrating HTML parsing, optical character recognition, and a large language model. Structured case profiles captured demographics, diagnostic history, barriers, facilitators, and patient-reported outcomes, with quality control applied. A blinded human review was conducted for preliminary validation. Logistic regression examined factors associated with favorable outcomes, defined as diagnostic delay of ≤ 2 months and absence of severe harms.

Results: Among 1656 unique cases, the median age was 47 years; 28.9% reported female and 32.8% male; 53% of posts were authored by family members. Psychological distress was reported in 88.5%, disease progression in 37.2%, and treatment-related harm in 16.6%. Unfavorable outcomes were strongly associated with clinician-related issues (68.5%; OR 0.26, 0.16-0.42), case complexity (67.4%; OR 0.30, 0.19-0.48), specialist access barriers (36.4%; OR 0.48, 0.30-0.76), and ≥3 diagnostic encounters (OR 0.44, 95% CI 0.19-1.00). Specialist input (60.0%; OR 2.02, 1.34-3.07) was the strongest facilitator, while patient self-advocacy (12.9%; OR 0.55, 0.33-0.93) and family support (12.0%; OR 0.54, 0.30-0.96) were more common in difficult cases. A patient journey map illustrated how barriers and facilitators interacted across care stages.

Conclusion: This large-scale patient narrative analysis highlights clinician-related issues, case complexity, and specialist access barriers as major contributors to patient-reported diagnostic challenges, while specialist input facilitates resolution. AI-driven text mining of online forums enables scalable, patient-centred insights into diagnostic pathways and points to opportunities for digital tools to improve overall patient experience.

Keywords: Lymphoma, Large Language Models, Patient Safety, Diagnostic Error, Patient-Reported Outcomes

INTRODUCTION

Lymphoma accounts for 3-4% of all cancers worldwide, with more than half a million new cases reported annually [1]. Accurate diagnosis of lymphoma is a major challenge in clinical practice, mainly because of its diverse subtypes and nonspecific symptoms in the early stages [2, 3]. Expert reviews in the United States, United Kingdom, and Europe have reported diagnostic discrepancies in 5.8-27.3% of the cases [2, 4, 5], while studies in China documented rates as high as 39.8% [6]. Such errors can delay appropriate care, expose patients to harmful treatment [4], and may contribute to survival disparities, with 5-year relative survival at 38.3% in China compared with over 70% in Europe [7, 8].

Most prior research has focused on preventing diagnostic errors through clinician training, advanced imaging, or pathology support [4, 5, 9-11], but little is known about how patients themselves navigate diagnostic discrepancies. Correcting errors often requires patients and families to advocate for second opinions, pursue tertiary care, and endure financial and psychological strain [9, 12], as overturning an initial diagnosis can be difficult for clinicians and institutional support is often lacking.

Patient narratives provide critical insights into these processes, but conventional surveys and registries capture them only at small scale and limited depth [13-15]. Recent advances in artificial intelligence, particularly large language models (LLMs), enable scalable and systematic analyses of unstructured text [16-18]. Online health communities host rich repositories of patient-reported experiences, yet they remain underexplored in current research. Applying LLMs to these narratives may therefore help identify factors associated with patient-perceived diagnostic challenges, providing a patient-centred perspective that complements existing clinical literature [15, 16, 19].

Building on this gap, we developed an AI-based pipeline to analyse forum posts relevant to misdiagnosis in a Chinese lymphoma community. Our objectives were to (1) characterize barriers and facilitators influencing the correction of diagnostic errors, (2) map these factors onto key stages of the patient journey, (3) quantify associations between these factors and patient-perceived impacts, and (4) explore the potential role of AI tools in mitigating diagnostic challenges and improving patient experience.

METHODS

Study Design and Data Collection

We conducted a retrospective cross-sectional study of patient narratives from House086.com, the largest Chinese online forum dedicated to lymphoma patients and caregivers (≥100000 registered members). Posts published between 2011 and 2025 that contained terms related to diagnostic error (e.g., misdiagnosis, diagnostic delay, second opinion, pathological discrepancy; full list in **Supplementary Table S1**) were identified through keyword searches in Chinese and manually reviewed to generate a curated list of candidate URLs at the time of data collection (June-August, 2025).

We developed an AI-based extraction pipeline (**Figure 1**) to retrieve raw HTML content and process attached images with optical character recognition (OCR) to capture clinical details from uploaded medical reports and supporting documents. Text and OCR-derived content were analysed by DeepSeek [20], an LLM with strong performance in Chinese text comprehension. Using structured prompts, the model identified cases with potential discrepancies, reconstructed diagnostic histories, selected associated factors and patient-reported impacts based on a predefined taxonomy, and assigned a case quality score (1-5). The pipeline then standardized outputs into structured JSON case profiles.

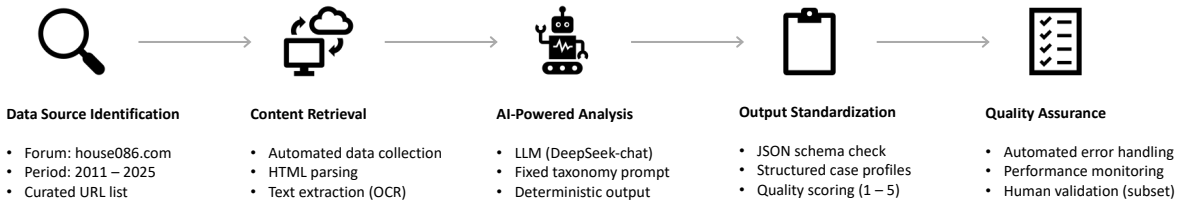


Figure 1. AI Pipeline for Extracting Cases from Patient Forum Posts.
The pipeline begins with a curated list of forum URLs, from which raw HTML and images were retrieved. Texts and contents from optical character recognition (OCR) were processed by a large language model to identify diagnostic discrepancies, reconstruct diagnostic histories, and classify barriers, facilitators, and impacts. Outputs were standardized into structured case profiles for statistical analysis.

Variable Definitions

Username, user type (patient, family member, other), and posting year were directly parsed from the HTML content, whereas patient profiles (age, sex, diagnostic history, barriers and

facilitators in the treatment journey, and impacts) were inferred by the LLM from narrative text.

Given the heterogeneity in content and format of patient narratives, the LLM was allowed to freely propose top keywords for barriers, facilitators, and impacts in pilot analyses. These keywords were clustered using latent Dirichlet allocation (LDA [21]) and labelled manually to develop a fixed taxonomy for balanced data representation and clinical interpretability, including 11 barriers (e.g., case complexity, specimen issues, clinician related issues), 7 facilitators (e.g., specialist input, tertiary hospitals, family support), and 5 impacts (disease progression, treatment-related harm, psychological distress, loss of time, and financial strain, see **Supplementary Table S2** for details). In the formal analyses, the model could assign up to three items per category for each case. The LLM also rated patient-reported impact severity and case quality on a 5-point scale. A case was classified as a diagnostic discrepancy when the narrative provided sufficient evidence of a diagnosis that was later contradicted or revised, as determined by the LLM from context rather than by fixed rules.

The primary outcome for quantitative analyses, "favourable outcome", was defined as both moderate-or-lower severity (score≤3) and diagnostic delay ≤ 2 months (the cohort median). "Unfavourable outcome" was defined as either severe impacts (≥4) or delays > 2 months, to account for both the biological aggressiveness in certain subtypes and the systemic burden of diagnostic delay, particularly in indolent forms. Diagnostic encounters were defined as distinct consultations or diagnostic conclusions reported in the patient's narrative, which could occur in the same or different hospitals.

Data Preprocessing

The workflow of case screening and inclusion is summarized in **Figure 2.** From over three million forum posts, 12496 were returned by keyword search. After manual review, 10462 were excluded as irrelevant to the study scope or duplicates of already selected posts, leaving 2034 URLs for processing. During AI-assisted analysis, 18 URLs failed content retrieval, whereas 2016 threads contributed 2406 narratives related to diagnostic discrepancies. Of these, narratives were excluded if they did not provide sufficient detail to support classification (n=449), had low case quality (n=9), or lacked meaningful description of patient-perceived impacts (n=21). For duplicate narratives from the same user, only the most

recent and highest-quality version was retained (excluding n=271). The final analytic dataset comprised 1656 unique case profiles for analyses.

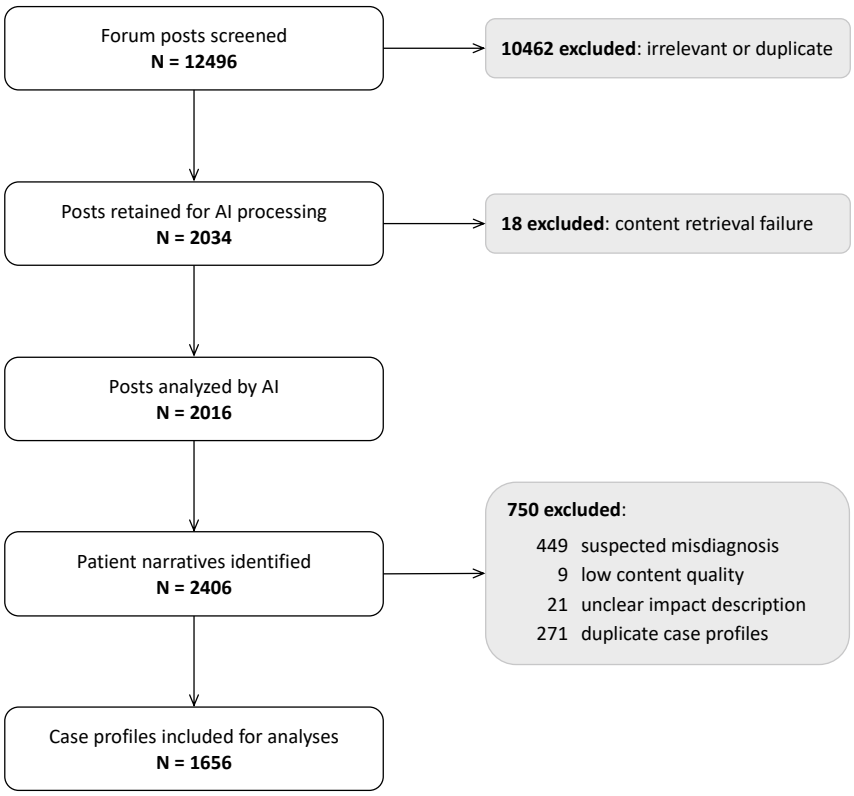


Figure 2. Data Cleaning Workflow for Patient Forum Posts.
From more than three million posts, irrelevant entries were excluded using keyword screening and manual review. Candidate posts were processed through the AI pipeline to extract diagnosis-related details. After quality control and de-duplication, 1656 unique case profiles were retained for analysis.

Due to the anonymous nature of forum discussions, 66.8% of the posts did not disclose patient age and 38.3% did not disclose sex. To retain sample size, these were coded as a separate category ("Not Disclosed"). Other missing covariates were handled by complete case analysis. The primary outcome (favorable vs. unfavorable) was available for 1059 narratives (63.9% of the analytic dataset) and missing in 597 (36.1%), mainly because delay duration was not reported.

Statistical Analyses

Categorical variables were summarized as counts (%) and continuous variables as medians with interquartile ranges (IQRs). Impact severity was reported as mean (SD) due to limited

variability in median values. Inter-group differences were tested with chi-square or t-tests as appropriate.

Association between determinants and the outcome were estimated using multivariable logistic regression, adjusting for predefined barriers, facilitators, age, sex, posting year, user type, and the number of diagnostic encounters. Age and posting year were categorized to account for potential non-linear trends. Odds ratios (ORs) with 95% confidence intervals (CIs) were reported. Multicollinearity among categorical variables was assessed using the phi coefficient. Analyses were conducted on all available data; cases with missing outcomes contributed to descriptive analysis but not regression modelling.

Sensitivity Analysis

To preliminarily assess pipeline reliability, six randomly selected cases were manually reviewed across all key extracted variables, including diagnostic history and impacts. Agreement exceeded 80%, with minor discrepancies limited to hospital names. Formal validation with larger samples and external medical review are planned as future work.

Ethics

This study analysed patient narratives from an open-access online forum. All data were publicly accessible without registration or login requirements. No identifiable personal information was collected or analysed. In accordance with applicable regulations and policies, ethics approval and patient consent were not required. The study adhered to the principles of the Declaration of Helsinki.

RESULTS

Study Population Characteristics

Of 1656 cases, the primary outcome was available for 1095 patients. Among these, 321 (29.3%) achieved favourable outcomes whereas 738 (67.4%) experienced unfavourable outcomes (**Table 1)**. Among the subset with reported age, unfavorable outcomes were more commonly observed in children and older adults, although interpretation was limited by the large proportion of missing data. Patients with unfavorable outcomes were more likely to have ≥3 diagnostic encounters compared with those achieving favourable outcomes. Median delay was 1.0 month [IQR 0.5-1.0] in favourable outcomes versus 4.5 months [2.0-12.0] in

unfavourable outcomes. Impact severity scores clustered at 3 (59.6%), with 34.7% as 4, and 1.6% as 5.

Table 1. Population characteristics by the patient-reported outcome

	Overall (N=1656)	Favourable outcome (n=321)	Unfavourable outcome (n=738)	p-value
Patient age, years	47.0 [30.0, 59.0]	49.0 [32.0, 59.0]	42.0 [27.0, 58.0]	0.085
Patient age category				<0.001
<18	45 (2.7)	6 (1.9)	33 (4.5)	
18-40	193 (11.7)	61 (19.0)	100 (13.6)	
40-60	180 (10.9)	68 (21.2)	81 (11.0)	
>=60	131 (7.9)	39 (12.1)	65 (8.8)	
Not disclosed	1107 (66.8)	147 (45.8)	459 (62.2)	
Patient sex				0.641
Female	479 (28.9)	108 (33.6)	244 (33.1)	
Male	543 (32.8)	130 (40.5)	283 (38.3)	
Not disclosed	634 (38.3)	83 (25.9)	211 (28.6)	
Post user type				0.863
Patient-self	765 (46.2)	137 (42.7)	325 (44.0)	
Family member	871 (52.6)	180 (56.1)	402 (54.5)	
Others (admin, doctors)	20 (1.2)	4 (1.2)	11 (1.5)	
Posting year				0.061
2011 – 2015	161 (9.7)	26 (8.1)	89 (12.1)	
2016 – 2020	602 (36.4)	101 (31.5)	253 (34.3)	
2021 – 2025	893 (53.9)	194 (60.4)	396 (53.7)	
No. of diagnostic encounters	2.0 [2.0, 3.0]	2.0 [2.0, 3.0]	2.0 [2.0, 3.0]	0.005
Category of diagnostic encounters				0.004
2	967 (61.6)	207 (65.5)	405 (56.8)	
>= 3	426 (27.2)	91 (28.8)	280 (39.3)	
Not disclosed	176 (11.2)	18 (5.7)	28 (3.9)	
No. of months delayed*	2.0 [1.0, 6.0]	1.0 [0.5, 1.0]	4.5 [2.0, 12.0]	<0.001
Impact severity score^	3.3 (0.6)	3.0 (0.2)	3.8 (0.5)	<0.001
Patient-reported impacts†				
Psychological distress	1466 (88.5)	301 (93.8)	663 (89.8)	0.052
Loss of time	781 (47.2)	195 (60.7)	429 (58.1)	0.467
Disease progression	616 (37.2)	43 (13.4)	512 (69.4)	<0.001
Treatment-related harm	275 (16.6)	23 (7.2)	182 (24.7)	<0.001
Financial strain	77 (4.6)	4 (1.2)	57 (7.7)	<0.001

* A small subset reported very long delays (>6 years, n=11; >10 years, n=4).
^ Number of diagnostic encounters and impact severity score reported as mean (SD) due to limited variability in median values.
† Multiple impacts could be reported by a single user.

Psychological distress was the most common impact (88.5%), followed by loss of time (47.2%), disease progression (37.2%), treatment-related harm (16.6%), and financial strain (7.6%). Disease progression and treatment-related harm were both significantly more frequent in the unfavorable outcomes (69.4% vs 13.4%, and 24.7% vs. 7.2%, respectively).

Patient Journey Map

Qualitative synthesis of narratives revealed a recurring sequence of four phases leading to treatment initiation or withdrawal (**Figure 3**): symptom onset and initial help-seeking within the family domain; the first medical encounter, where diagnostic problems were most often reported; patient decision-making after receiving an initial diagnosis; and subsequent encounters, often involving second opinions and sometimes multiple rounds before consensus. Treatment initiation generally followed these phases, although some patients described seeking further re-diagnosis even after starting therapy. Each phase showed distinct patterns of barriers and facilitators, and impacts in early phases often created new barriers in later stages, compounding the difficulty of correction. Some factors, such as patient self-advocacy and family support, were relevant across multiple phases, whereas others were phase-specific (e.g., specimen issues in phases 2 and 4).

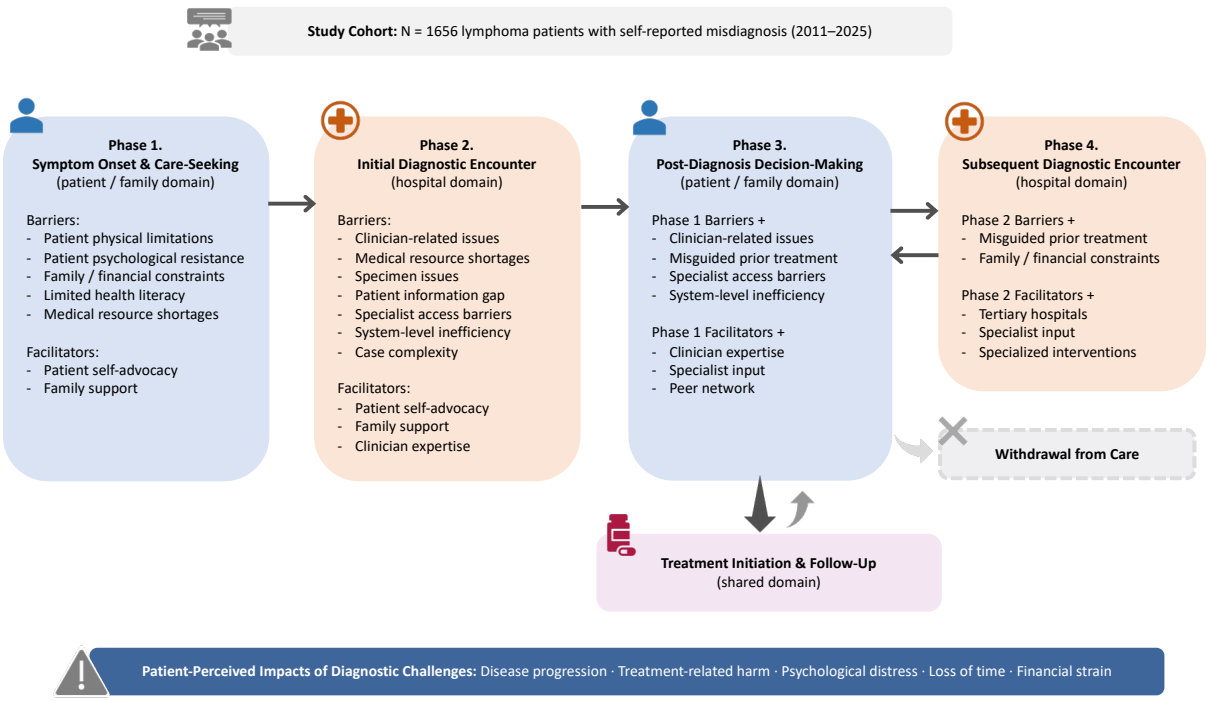


Figure 3. Patient Journey Map of Forum Cases.

The map illustrates four recurring phases from symptom onset to treatment initiation. Barriers and facilitators identified from forum narratives are shown within each phase. Patient-perceived impacts (disease progression, loss of time, financial strain, psychological distress, and treatment-related harm) are summarized at the bottom. Withdrawal from care is displayed as an alternative pathway, and some patients described seeking further re-diagnosis even after starting therapy, reflecting the non-linear nature of the diagnostic journey.

Regression Results

In multivariable models (**Figure 4**, with data in **Table S3)**, clinician-related issues (68.5%; OR 0.26, 95% CI 0.16-0.42) and case complexity (67.4%; OR 0.30, 0.19-0.48) were the most prevalent barriers, both strongly associated with unfavorable outcomes. Specialist access barriers (36.4%), medical resource shortages (15.2%), and specimen issues (13.0%) were also significant predictors. Less frequent barriers, including patient psychological resistance, limited health literacy, and family / financial constraints (<6% each), were nonetheless associated with reduced odds of favourable outcome.

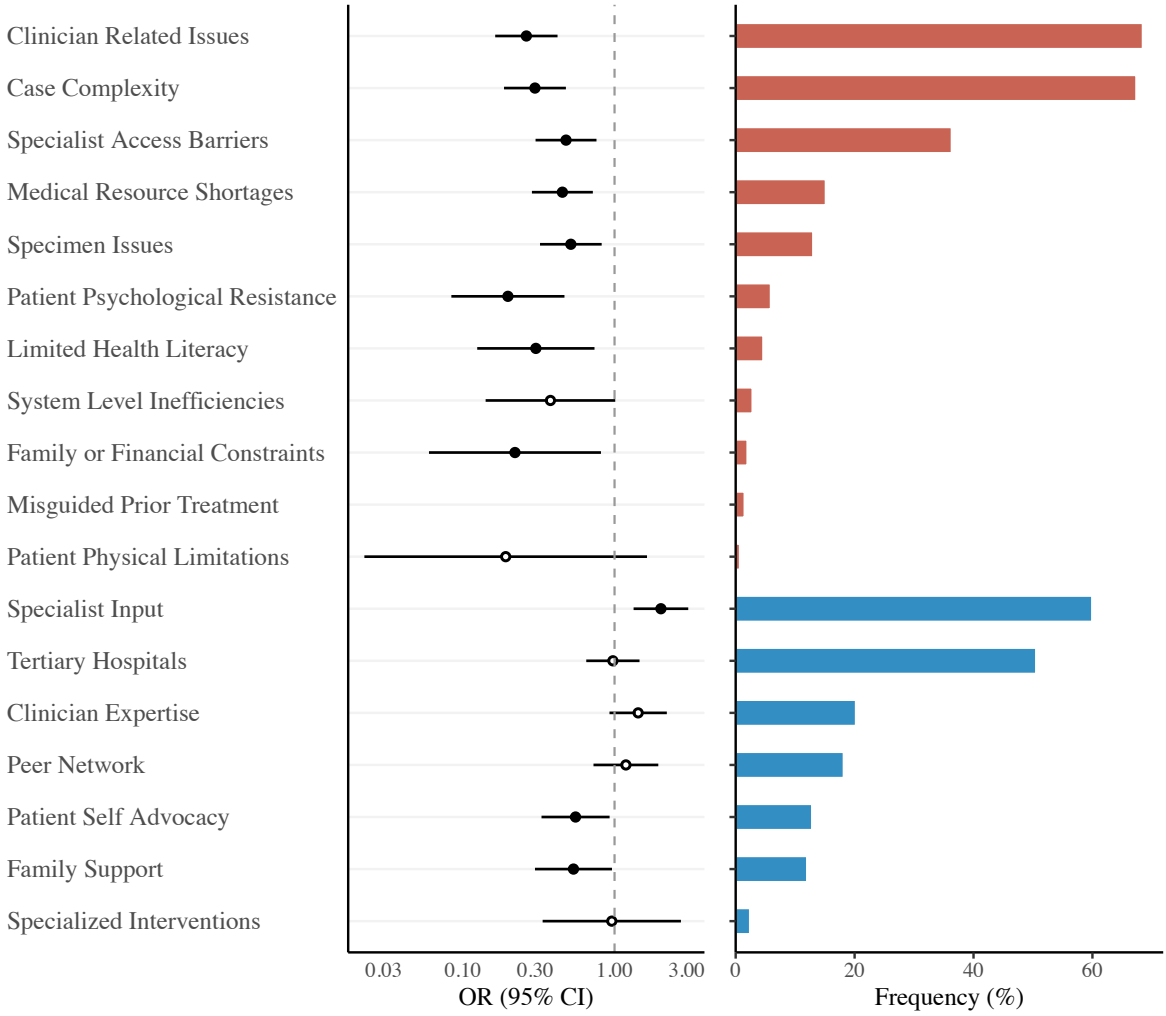


Figure 4. Barriers and Facilitators Associated With Patient-Reported Outcomes

Left: Forest plot from multivariable logistic regression showing odds ratios (95% CI) for each barrier and facilitator in relation to favourable outcome (defined as severity ≤3 and delay ≤2 months). Misguided prior treatment was excluded from the model due to perfect separation. Filled dots indicate $p \leq 0.05$; open dots indicate $p > 0.05$. Right: Frequency of barriers and facilitators across the analytic cohort (n = 1656). Bars show the proportion of patient narratives mentioning each factor, with barriers in red and facilitators in blue, ordered by descending frequency within each category.

Among facilitators, specialist input was most frequent (60.0%) and strongly associated with favourable outcomes (OR 2.02, 1.34-3.07). Patient self-advocacy (12.9%) and family support (12.0%) were also significant facilitators, whereas tertiary hospital access, clinician expertise, peer networks, and specialized interventions were not.

Having ≥ 3 diagnostic encounters was associated with lower odds of favourable outcomes (OR 0.44, 0.19-1.00). Although the number of posts increased chronologically, outcome distribution did not differ across posting years when grouped into 3-year or 5-year categories.

Collinearity testing showed moderate correlation between clinician-related issues and specialist access barriers (phi=0.55), but both were retained due to distinct clinical implications. Misguided prior treatment, defined as inappropriate or harmful interventions before an accurate diagnosis, was reported only in unfavorable outcomes, and was excluded from the model to avoid perfect separation.

Validation

In a small pilot check (n=6), over 80% alignment was observed between AI outputs and manual review across >10 extracted variables, with discrepancies mainly in hospital names.

DISCUSSION

Key Findings

This large-scale analysis of 1656 patient narratives identified both clinical and patient-level factors influencing the experience of lymphoma diagnosis. Clinician-related issues (68.5%) and case complexity (67.4%) were the most frequent barriers, while less common but significant barriers included patient psychological resistance, limited health literacy, and family or financial constraints. Specialist input (60.0%) emerged as the leading facilitator for favorable outcomes, whereas patient self-advocacy (12.9%) and family support (12.0%) appeared more often in complex or delayed cases, reflecting their role in navigating difficult diagnostic journeys. Finally, having ≥3 diagnostic encounters was linked to unfavorable outcomes, reflecting cumulative complexity.

Comparison with Existing Evidence

Our barrier-facilitator framework echoes prior work on cancer care quality, where delays, fragmentation, and inadequate psychosocial support were key barriers, and multidisciplinary

coordination and IT integration acted as facilitators [15]. Our study extends this literature by identifying challenges specific to lymphoma: The high prevalence of case complexity (67.4%) and specimen issues (13.0%) is consistent with longstanding recognition that pathology in lymphoma is particularly challenging [2, 4, 5]. Similarly, our findings that repeated diagnostic encounters predicted unfavorable outcomes align with survey data from UK, where lymphoma patients required more pre-referral consultations than those with most other cancers, reflecting intrinsic diagnostic complexity [3].

From the patient perspective, clinician-related issues, including inadequate communication, unprofessional attitudes, low diagnostic accuracy, and diagnostic errors were the most frequently cited barrier (68.5%). While this does not necessarily imply malpractice, it highlights gaps in training and communication that compromise quality care [9, 22]. Inappropriate initial treatment, although rare (1.5%), occurred exclusively in the unfavourable group, exemplifying the risks of initiating therapy without diagnostic certainty. Supporting this, Proctor et al. reported that 11% of discordant lymphoma diagnoses required major change in treatment [4]. Even when direct clinical harm was minimal, patients faced substantial financial burden from referral, ancillary testing, and unnecessary treatment [4, 9]. Such systemic vulnerabilities may contribute to the observed survival gap, with 5-year survival at 38.3% in China compared with over 70% in Europe [7, 8].

Interestingly, patient self-advocacy and family support were negatively associated with favorable outcomes. Beyond possible retrospective narrative bias, this may reflect confounding by case complexity or delayed diagnosis, where such behaviors emerge as “last resort” strategies after prolonged uncertainty [3, 9]. Prior literature also describes physician defensiveness toward assertive patients (“difficult patient” bias), which may further compromise care [23, 24]. These findings suggest that individual-level efforts are insufficient without systemic support, in contrast to specialist input that represents structured intervention. The predominance of family-authored narratives (>50%) highlights the central role of family support in advocating for diagnostic review, particularly for children and the elderly, who may have limited capacity for self-advocacy.

Psychological distress was the most frequently reported impact (88.5%), followed by loss of time (47.2%). While not directly predictive of outcomes, distress can evolve into resistance to further diagnostic testing, reinforcing barriers to correction. Financial strain, reported more

often in unfavourable outcomes (7.7% VS. 1.2%), reflects household-level barriers to sustained care [4]. Clinically relevant harms, including disease progression (69.4% VS. 13.4%) and treatment-related harm (24.7% VS. 7.2%), were also markedly more common in unfavorable outcomes, increasing complexity and misguiding subsequent therapy. Similar to our taxonomy, Mazor et al. described psychological, family, and financial consequences of problematic care events, revealing the multifaceted nature of patient-perceived harm [12]. As illustrated in the patient journey map (**Figure 3**), the diagnosis of lymphoma is not a single-point challenge, but a multi-phase, multi-factor process, where early impacts can evolve into new barriers, underscoring the need for coordinated efforts across clinicians, patients, families, and health systems [9, 14].

Implication for Practice and Digital Health Applications

Our findings suggest that targeted interventions are needed to systematically address barriers to patient-centred lymphoma care. International initiatives such as the Nebraska Lymphoma Study Group, the North Central London Lymphoma network, and the French Lymphopath Network show how centralized expert review can reduce diagnostic errors [2, 4, 5]. However, such frameworks remain difficult to implement in countries with limited resources and large patient populations. In China, for example, patients are often responsible for arranging the transfer of pathology slides themselves, reflecting gaps in referral infrastructure.

Digital health tools may provide complementary solutions at multiple levels. First, AI-assisted pathology has achieved high accuracy in lymphoma detection and could reduce specimen-related errors and case complexity [10, 11]. Second, patient-facing LLM applications such as chatbots may enhance health literacy, provide psychological support, and facilitate shared decision-making, potentially easing strained physician-patient relationships and fostering trust [17, 18, 22]. Third, system-level tools such as decision-support platforms, triage algorithms, telemedicine, and outpatient robotics could help address financial, logistical, and physical barriers, particularly in under-resourced regions.

Patients are already adopting AI tools independently to interpret pathology reports and disease information, as observed during manual screening. Recent work has raised concerns that patients may over-trust AI-generated advice even when inaccurate [25]. Our findings complement this perspective by examining a population already exposed to potential diagnostic inaccuracies, often in settings where specialist access is limited. In such contexts,

patients described turning to AI for its perceived neutrality and convenience, underscoring the urgent need for health systems to integrate safe, evidence-based digital tools with appropriate medical oversight.

Strengths and Limitations

This study has several strengths. To our knowledge, it is the first large-scale analysis focusing specifically on patients potentially exposed to misdiagnosis, an important but often overlooked population in oncology care. By applying large language models to online patient narratives, we captured authentic perspectives less constrained by institutional framing, and highlighted the value of patient voices in chronic disease management [16]. The AI pipeline provided a consistent and scalable method for structuring unstructured text, enabling rapid analysis of thousands of narratives and offering a transferable framework to other diseases and settings. Unlike prior malpractice- or insurance-based studies using standardized harm severity codes [26], our outcome definition was derived from patient-perceived impacts, offering complementary insights into care-seeking behaviours and psychological burden.

Several limitations must be acknowledged: First, missing demographic data (e.g., age, sex) limited subgroup analyses. Second, OCR occasionally failed on extracting texts from older low-resolution images. A predefined limit of reading up to 10 images per post may have introduced information loss, which was partly mitigated by LLM inference and case quality scoring. Finally, although a preliminary validation showed >80% alignment between AI outputs and manual review in six cases, more robust evaluation with larger sample sizes, formal κ statistics, and external expert review are needed to confirm accuracy and generalizability.

Future Research

Future work will first focus on systematic human validation to ensure the robustness and accuracy of the AI pipeline. A random sample of 100 posts will be double-annotated by two bilingual medical researchers, blinded to AI outputs, with inter-rater reliability assessed using Cohen's κ and discrepancies resolved by consensus. Further prospective validation against clinical records could strengthen reliability. In addition, longitudinal analyses of repeated narratives may reveal the dynamic interplay of barriers and facilitators over time, distinguishing short- from long-term impacts on patients and families. Health economics methods could also provide a more comprehensive assessment of the societal burden of

diagnostic errors. Beyond lymphoma, the analytic framework could be extended to patient-authored texts in other diseases and healthcare systems, enabling systematic incorporation of patient perspectives into clinical research and practice.

CONCLUSION

In summary, this large-scale, LLM-based analysis of more than 1,600 forum narratives identified key barriers and facilitators shaping lymphoma patient experiences after diagnostic errors. The findings highlight that effective correction requires not only hospital-based interventions but also active engagement of patients, families, peer networks, and health systems. More broadly, this study demonstrates the feasibility of using patient-generated data to inform clinical practice and digital health innovation, underscoring the potential of AI tools to strengthen diagnostic pathways and improve patient-perceived quality of care.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ACKNOWLEDGMENTS

We thank participants of online patient forums for sharing their experiences, which made this research possible. ChatGPT was used for language polishing, and Claude was used for code development. All scientific content and analyses were conducted and verified by the author.

DATA & CODE AVAILABILITY

All codes used for data collection, preprocessing, and statistical analysis will be made publicly available upon publication. The de-identified dataset may be shared upon reasonable request.

REFERENCES

REFERENCES

1. Sung, H., et al., *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: a cancer journal for clinicians, 2021. **71**(3): p. 209-249.
2. Laurent, C., et al., *Impact of expert pathologic review of lymphoma diagnosis: study of patients from the French Lymphopath Network.* Journal of clinical Oncology, 2017. **35**(18): p. 2008-2017.
3. Lyratzopoulos, G., et al., *Variation in number of general practitioner consultations before hospital referral for cancer: findings from the 2010 National Cancer Patient Experience Survey in England.* The Lancet Oncology, 2012. **13**(4): p. 353-365.
4. Proctor, I.E., et al., *Importance of expert central review in the diagnosis of lymphoid malignancies in a regional cancer network.* Journal of Clinical Oncology, 2011. **29**(11): p. 1431-1435.
5. Bowen, J.M., et al., *Lymphoma diagnosis at an academic centre: rate of revision and impact on patient care.* Br J Haematol, 2014. **166**(2): p. 202-8.

6. Deng, J., et al., *Misdiagnosis analysis of 2291 cases of haematolymphoid neoplasms.* Frontiers in Oncology, 2023. **13**: p. 1128636.
7. Liu, W., et al., *Mortality of lymphoma and myeloma in China, 2004–2017: an observational study.* Journal of Hematology & Oncology, 2019. **12**(1): p. 22.
8. Sant, M., et al., *Survival for haematological malignancies in Europe between 1997 and 2008 by region and age: results of EUROCARE-5, a population-based study.* The Lancet Oncology, 2014. **15**(9): p. 931-942.
9. Singh, H., et al., *The global burden of diagnostic errors in primary care.* BMJ Qual Saf, 2017. **26**(6): p. 484-494.
10. Syrykh, C., et al., *Accurate diagnosis of lymphoma on whole-slide histopathology images using deep learning.* NPJ digital medicine, 2020. **3**(1): p. 63.
11. Li, D., et al., *A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals.* Nature communications, 2020. **11**(1): p. 6004.
12. Mazor, K.M., et al., *Toward patient-centered cancer care: patient perceptions of problematic events, impact, and response.* Journal of Clinical Oncology, 2012. **30**(15): p. 1784-1790.
13. Platt, J.E., P.D. Jacobson, and S.L.R. Kardia, *Public Trust in Health Information Sharing: A Measure of System Trust.* Health Serv Res, 2018. **53**(2): p. 824-845.
14. Foglino, S., et al., *The relationship between integrated care and cancer patient experience: A scoping review of the evidence.* Health Policy, 2016. **120**(1): p. 55-63.
15. Wagner, E.H., et al., *The quality of cancer patient experience: perspectives of patients, family members, providers and experts.* Quality and Safety in Health Care, 2010. **19**(6): p. 484.
16. Dreisbach, C., et al., *A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data.* Int J Med Inform, 2019. **125**: p. 37-46.
17. Arora, A. and A. Arora, *The promise of large language models in health care.* The Lancet, 2023. **401**(10377): p. 641.
18. Thirunavukarasu, A.J., et al., *Large language models in medicine.* Nature Medicine, 2023. **29**(8): p. 1930-1940.

19. Siddiqui, Z.A., et al., *Leveraging social media data to study disease and treatment characteristics of Hodgkin's lymphoma Using Natural Language Processing methods.* PLOS Digit Health, 2025. **4**(3): p. e0000765.
20. Gibney, E., *China's cheap, open AI model DeepSeek thrills scientists.* Nature, 2025. **638**(8049): p. 13-14.
21. Walsh, J., J. Cave, and F. Griffiths, *Combining Topic Modeling, Sentiment Analysis, and Corpus Linguistics to Analyze Unstructured Web-Based Patient Experience Data: Case Study of Modafinil Experiences.* J Med Internet Res, 2024. **26**: p. e54321.
22. Salwei, M.E., J.S. Ancker, and M.B. Weinger, *The decision aid is the easy part: workflow challenges of shared decision making in cancer care.* JNCI: Journal of the National Cancer Institute, 2023. **115**(11): p. 1271-1277.
23. Groves, J.E., *Taking care of the hateful patient.* N Engl J Med, 1978. **298**(16): p. 883-7.
24. Hahn, S.R., et al., *The difficult doctor-patient relationship: somatization, personality and psychopathology.* Journal of clinical epidemiology, 1994. **47**(6): p. 647-657.
25. Shekar, S., et al., *People Overtrust AI-Generated Medical Advice despite Low Accuracy.* NEJM AI, 2025. **2**(6): p. AIoa2300015.
26. Newman-Toker, D.E., et al., *Serious misdiagnosis-related harms in malpractice claims: The "Big Three" – vascular events, infections, and cancers.* Diagnosis, 2019. **6**(3): p. 227-240.

Data Source Identification

- Forum: house086.com
- Period: 2011 – 2025
- Curated URL list

Content Retrieval

- Automated data collection
- HTML parsing
- Text extraction (OCR)

AI-Powered Analysis

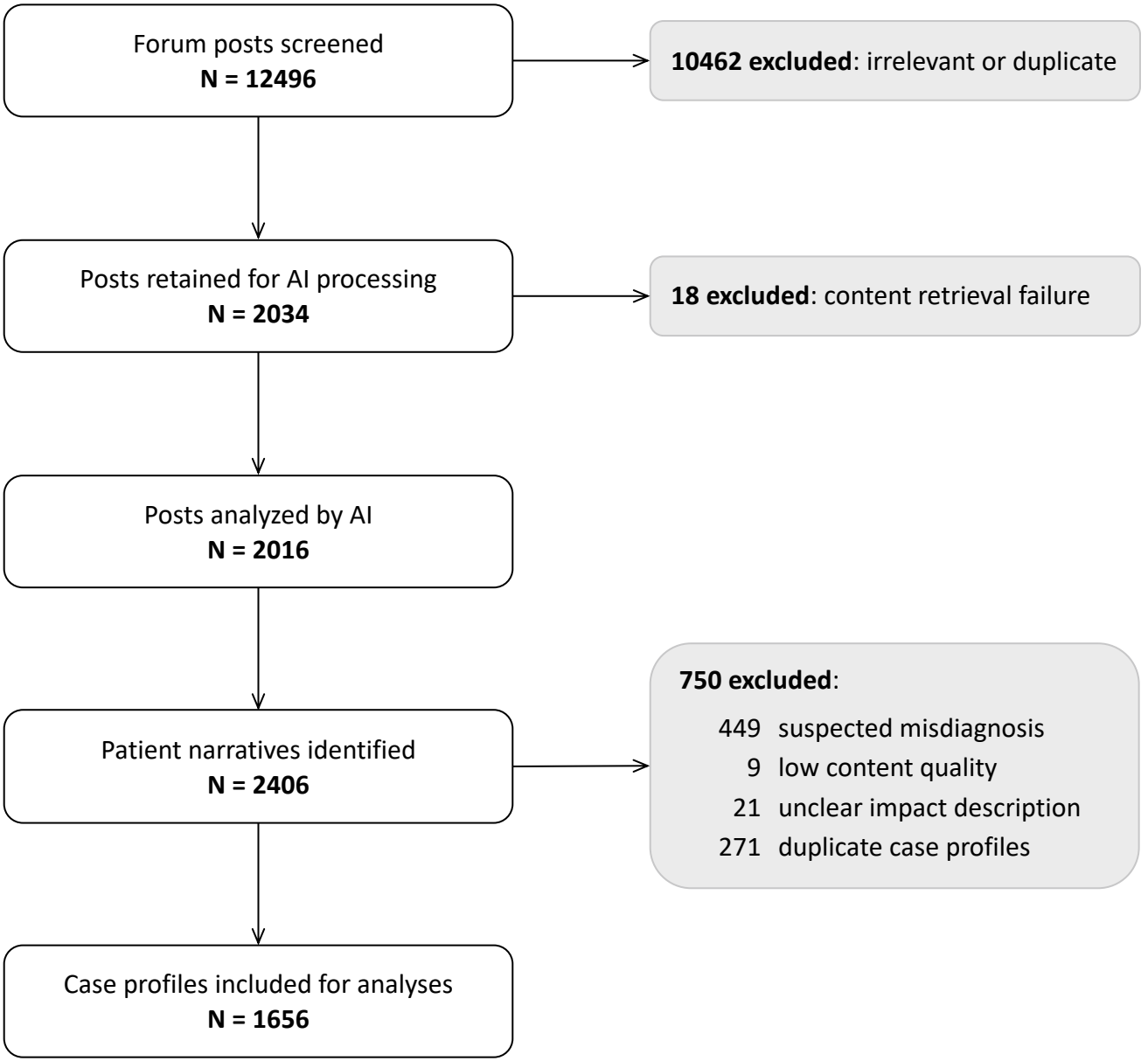
- LLM (DeepSeek-chat)
- Fixed taxonomy prompt
- Deterministic output

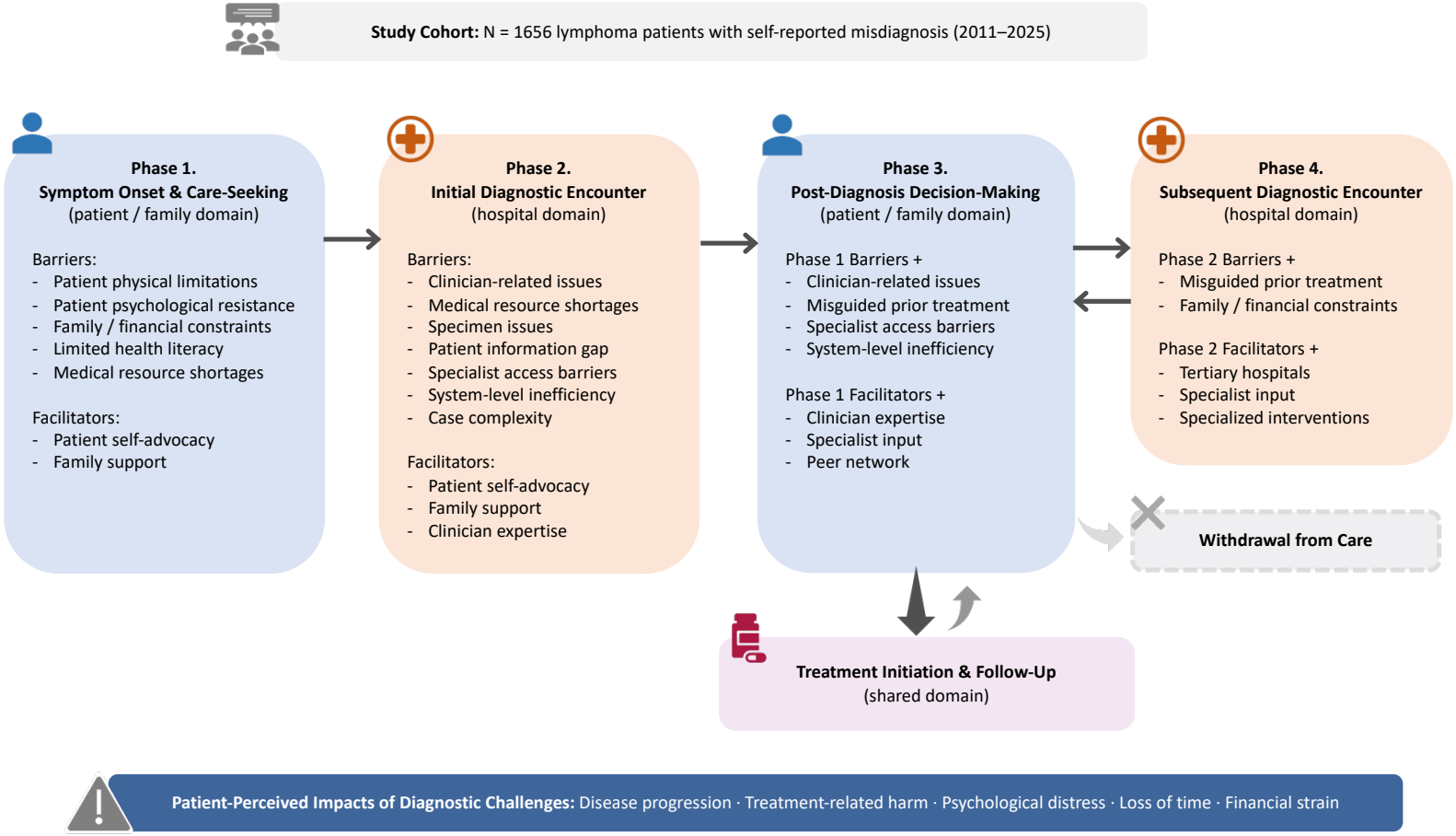
Output Standardization

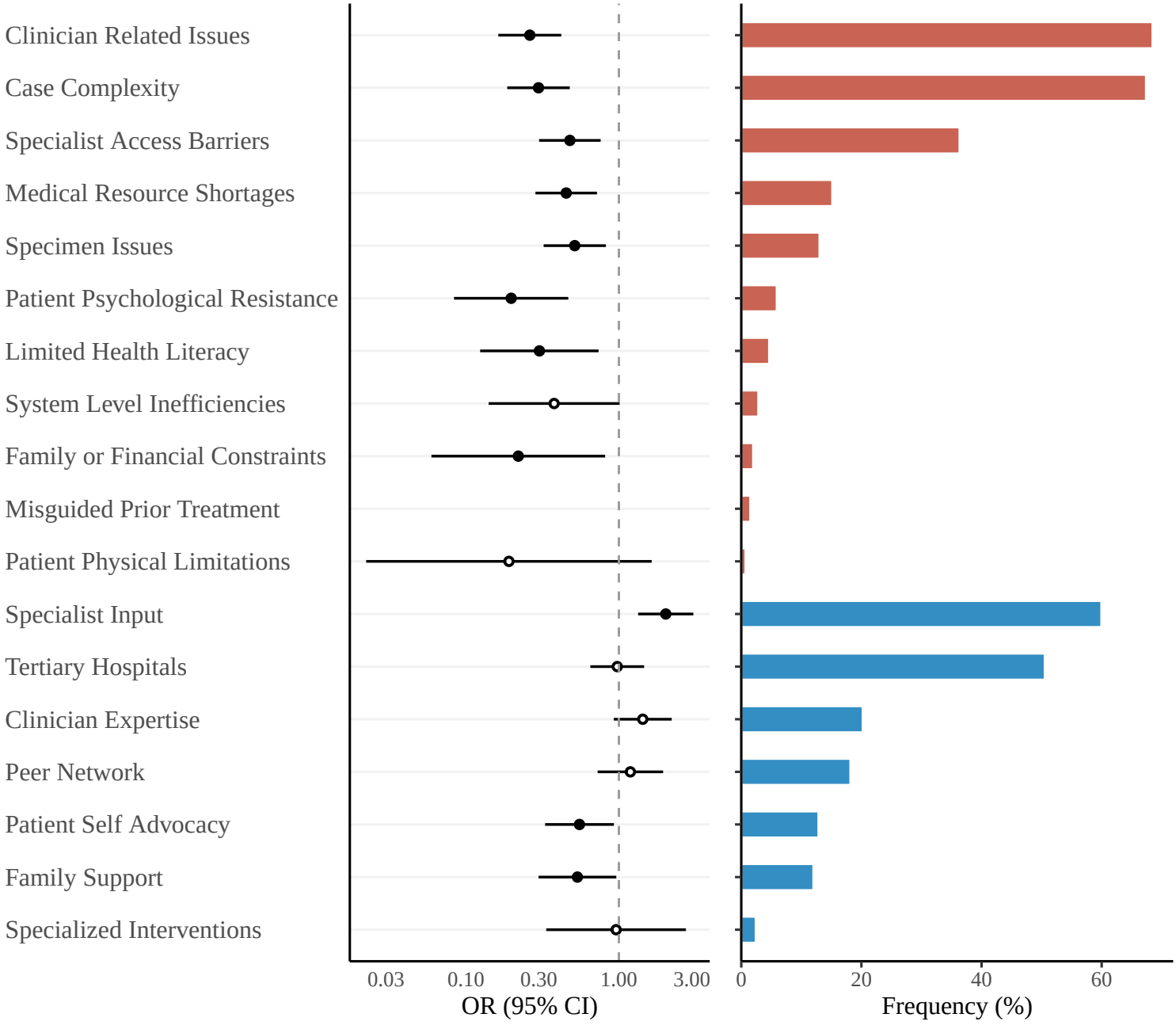
- JSON schema check
- Structured case profiles
- Quality scoring (1 – 5)

Quality Assurance

- Automated error handling
- Performance monitoring
- Human validation (subset)


Forum posts screened
N = 12496
10462 excluded: irrelevant or duplicate
Posts retained for AI processing
N = 2034
18 excluded: content retrieval failure
Posts analyzed by AI
N = 2016
Patient narratives identified
N = 2406
750 excluded:
449 suspected misdiagnosis
9 low content quality
21 unclear impact description
271 duplicate case profiles
Case profiles included for analyses
N = 1656


Study Cohort: N = 1656 lymphoma patients with self-reported misdiagnosis (2011–2025)
Phase 1.
Symptom Onset & Care-Seeking
(patient / family domain)
Barriers:
- Patient physical limitations
- Patient psychological resistance
- Family / financial constraints
- Limited health literacy
- Medical resource shortages
Facilitators:
- Patient self-advocacy
- Family support
Phase 2.
Initial Diagnostic Encounter
(hospital domain)
Barriers:
- Clinician-related issues
- Medical resource shortages
- Specimen issues
- Patient information gap
- Specialist access barriers
- System-level inefficiency
- Case complexity
Facilitators:
- Patient self-advocacy
- Family support
- Clinician expertise
Phase 3.
Post-Diagnosis Decision-Making
(patient / family domain)
Phase 1 Barriers +
- Clinician-related issues
- Misguided prior treatment
- Specialist access barriers
- System-level inefficiency
Phase 1 Facilitators +
- Clinician expertise
- Specialist input
- Peer network
Phase 4.
Subsequent Diagnostic Encounter
(hospital domain)
Phase 2 Barriers +
- Misguided prior treatment
- Family / financial constraints
Phase 2 Facilitators +
- Tertiary hospitals
- Specialist input
- Specialized interventions
Withdrawal from Care
Treatment Initiation & Follow-Up
(shared domain)
Patient-Perceived Impacts of Diagnostic Challenges: Disease progression · Treatment-related harm · Psychological distress · Loss of time · Financial strain


Clinician Related Issues
Case Complexity
Specialist Access Barriers
Medical Resource Shortages
Specimen Issues
Patient Psychological Resistance
Limited Health Literacy
System Level Inefficiencies
Family or Financial Constraints
Misguided Prior Treatment
Patient Physical Limitations
Specialist Input
Tertiary Hospitals
Clinician Expertise
Peer Network
Patient Self Advocacy
Family Support
Specialized Interventions
0.03
0.10
0.30
1.00
3.00
OR (95% CI)
0
20
40
60
Frequency (%)