LLM を用いた入院患者カルテからの血液悪性腫瘍早期兆候検出

○ 川本章太 [1,2]　宮本太郎 [2]　園部真也 [2]
[1] 慶應義塾大学 大学院政策・メディア研究科
[2] 東北大学病院
{shota.kawamoto.b5, taro.miyamoto.d1, shinya.sonobe.d7}@tohoku.ac.jp

概要

血液悪性腫瘍は発熱や倦怠感といった非特異的症状から始まるため、骨髄像検査に至るまでの遅延が問題となる。本研究では、骨髄像検査を予定する入院患者 92 例（陽性 46/陰性 46）について、検査 30〜4 日前のカルテテキストを整備し、大規模言語モデル（LLM）による早期兆候検出の実現可能性を評価した。Local/Global/Temporal という Context 管理と Rule-based/Zero-shot/Reflection というプロンプト設計を組み合わせた 7 条件を比較し、侵襲的検査の意思決定を模した評価設定を構築した。Reflection Local 条件では感度 67%、特異度 72%を達成し、検査 2 週間前（Day -14）の時点で陽性患者の 50%を捕捉できた。

1　はじめに

血液悪性腫瘍（白血病、悪性リンパ腫、多発性骨髄腫など）は、早期発見・早期治療が予後改善に直結する疾患群である。しかし、初期症状は発熱、倦怠感、貧血など非特異的であり、確定診断に至るまでに時間を要することがある。骨髄像検査は血液悪性腫瘍の確定診断に不可欠な検査であるが、侵襲的であるため、臨床医が検査の必要性を判断するまでに一定の時間を要する。

近年、大規模言語モデル（LLM）の医療応用が注目されている。電子カルテに記録された診療情報から疾患の兆候を自動検出することで、臨床医の診断支援が期待される。本研究では、入院患者の電子カルテテキストから LLM を用いて血液悪性腫瘍の兆候を早期検出する手法を提案し、その有効性を評価する。

本研究では以下の Research Questions に取り組む：

RQ1　LLM は入院から骨髄像検査までのカルテから血液悪性腫瘍を予測できるか？

RQ2　Context 管理戦略（Local vs Global vs Temporal）による精度差はあるか？

RQ3　陰性群での偽陽性率はどの程度か？

英語圏では、入院記録からのゼロショットコーディングや危険度層別化に LLM を適用する実践研究がすでに進んでいる [1]。一方で、日本語臨床 NLP における公開リソースは実症例 100 件規模の Real-MedNLP[2] や国家試験ベースの JMedBench[3] に限られ、長期入院カルテのノイズや日常表現を含むデータセットが不足している。本研究で整備したコホートは、このデータデザートを補完し、ガイドライン準拠の意思決定を検証できる国内初の基盤となる。

2　関連研究

血液悪性腫瘍の診断遅延は長年の課題である。Abel ら [4] は非特異的症状への依存が遅延の主因と報告し、Koshiaris ら [5] は初診から診断まで中央値 163 日と示した。日本血液学会ガイドライン [6, 7] は骨髄穿刺の実施基準を定めているが、臨床現場での即時想起は難しく、AI 支援の重要性が高まっている。

電子カルテからの臨床情報抽出では、Gholipour ら [8] がルールベース手法の優位性を示し、Kehl ら [9] は RNN で病勢進行判定に AUROC 0.86–0.94 を達成した。近年、LLM の医療応用が進展し、Namasivayam ら [10] は 40 種類のがん診断抽出で F1 > 0.85 を達成した。しかし、日本語臨床 NLP の公開ベンチマークは Real-MedNLP[2] 等の小規模データに限られ、長期入院カルテの評価環境は不足している。

LLM による縦断的臨床データ処理では、D'souza ら [11] が時系列要約で AUROC 0.79 を達成した一

表 1　コホート定義

群	定義	候補	最終
陽性	検査後に血液悪性腫瘍診断	2,718	46
陰性	悪性腫瘍なし（前悪性・固形含む）	92	46
合計			**92**

方、初期誤判断の累積（スノーボール効果）[12] や LLM の自己修正限界 [13] が課題とされる。本研究は、入院患者の診療記録から LLM を用いて血液悪性腫瘍の兆候を検査前に検出する初の試みである。

3　方法

3.1　コホート定義

対象は東北大学病院において 2014 年 4 月から 2024 年 3 月の間に入院し、骨髄像検査を受けた患者（n=3,250）である。血液悪性腫瘍（ICD-10: C81–C96, C90）の診断有無により以下の 2 群に分類した。

陽性群は骨髄像検査後に血液悪性腫瘍と診断された患者（n=2,718）、陰性群は血液悪性腫瘍の診断がなく、かつ前悪性腫瘍（MDS 等：D45, D46, D47）および固形腫瘍（C00–C80）を有しない患者（n=92）とした。前悪性腫瘍・固形腫瘍を除外したのは、これらの疾患が血液悪性腫瘍と類似の臨床所見を呈する可能性があるためである。

さらに、評価期間（Day -30〜Day -4）にカルテテキストが存在する患者に限定した結果、陰性群は 46 名となった。陽性群は陰性群に合わせて 46 名をランダムサンプリングし、最終的なコホートは 92 名となった（表 1）。

評価対象となるカルテテキストは、骨髄像検査の 30 日前から 4 日前（Day -30〜Day -4）に記録されたプログレスノートおよび看護 SOAP とした。検査直前の 3 日間（Day -3〜Day 0）は、検査オーダーに関連する情報が含まれる可能性があるため除外した。

骨髄像検査日（Day 0）は検査実施記録から特定し、複数回入退院している患者も検査当日のエピソードに紐づけた。Day -30〜Day -4 の観察期間にカルテテキストが存在しない患者は除外し、陽性・陰性とも 46 例へマッチングした。本研究は東北大学病院倫理審査委員会の承認のもと実施した。

3.2　カルテテキスト抽出と前処理

電子カルテからはプログレスノート、入院時記録、看護 SOAP を抽出し、Day -30〜Day -4 の範囲で時系列に整列させた。各エントリには記録日時と診療部署を保持し、Day 0 までの残日数 `days_to_test` を付与して解析単位を統一した。

前処理では、患者識別情報を除去し、全角制御記号や連続空白を正規化して可読性を確保した。LLM に渡す際は古い記録から順に連結し、コンテキスト戦略で設定する最大長（後述）を超えないよう切り詰めた。

表 2　Day -30〜Day -4 テキストにおける患者背景統計（中央値 [四分位範囲]）

指標	陽性 (n=46)	陰性 (n=46)
記録件数/患者	4 [1, 14]	12 [2, 77]
最も早い記録日 (Day)	−20 [−28, −10]	−26 [−30, −16]
最も遅い記録日 (Day)	−8 [−13, −4]	−4 [−7, −4]
観察窓カバレッジ（日）	7 [0, 18]	14 [4, 25]
合計文字数/患者	1,239 [446, 4,673]	3,872 [640, 16,624]
Progress ノート割合	52.9%	50.2%
看護 SOAP 割合	47.1%	49.8%

3.3　患者背景

Day -30〜Day -4 の観察窓で十分なテキストが取得できた陽性 46 例・陰性 46 例について、患者あたり記録件数や記録種別の分布を表 2 に示す。陽性群は中央値 4 件と記録件数が少なく、観察窓のカバー範囲も 7 日と限定的であった一方、陰性群は 12 件（IQR 2–77）と記録が多く、最初の記録が Day -26 付近から開始していた。Progress ノートと看護 SOAP の比率は両群でほぼ同等であり、看護師記録も LLM 入力に寄与していることが分かる。年齢や性別などの個人属性は匿名化のため取得していない。

3.4　Context 管理戦略

LLM への入力コンテキストの構成方法として、以下の 3 戦略を比較した。

Local　各評価時点において、直近の数日分のテキストのみを入力する。過去の情報は含まれない。

Global　入院日から評価時点までの全テキストを累積的に入力する。

Temporal　前日までの状態要約（State）と当日のテキストを入力し、State を更新しながら時系列を追跡する。

Local 戦略ではチェックポイント集合 $\mathcal{C} = \{30, 21, 14, 7, 4\}$（Day -30, -21, -14, -7, -4 に相当）を定義し、各 $C \in \mathcal{C}$ の直前 7 日以内（Day -4 のみ 4 日

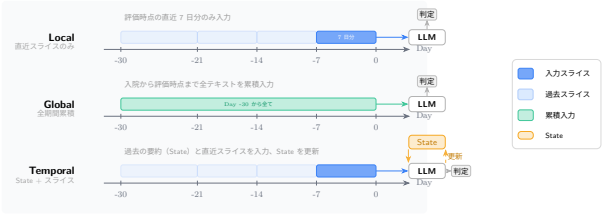

図 1　Context 管理戦略の比較。Local 戦略は評価日直近の 7 日分のみ、Global 戦略は全期間を累積的に、Temporal 戦略は過去の要約（State）と直近スライスを組み合わせて入力する。

間）のテキストのみを入力する。Global 戦略では Day -30 から対象チェックポイントまでの全レコードを結合し、Temporal 戦略は Local 戦略のスライスを逐次処理しながら State JSON を更新する。いずれの戦略でも、1 チャンクの最大長を 6,000 文字に制限し、カルテ概要と本文を保持したまま LLM の 8k トークン枠内へ収めた。

3.5　プロンプト設計

3 段階のプロンプト設計を比較した：

Lv1 Rule-based　血液学会ガイドラインに基づくキーワードリスト（発熱、盗汗、体重減少、リンパ節腫脹、LDH 上昇等）の有無を判定する。

Lv2 Zero-shot　LLM に対して「このカルテから血液悪性腫瘍を疑うべきか」を直接判断させる。

Lv3 Reflection　Zero-shot の判断に対して、「感染症や膠原病など他の可能性は？」と自問自答させ、最終判断を出力させる。

Rule-based 条件はキーワードの出現数に応じてリスクスコアを算出し（1 語 30 点、最大 100 点）、閾値 50 点以上またはキーワード検出で擬陽性とした。Zero-shot/Reflection 条件では医師ロールのシステムプロンプトを用い、`risk_score`（0–100）、`is_suspected`（真偽値）、`reasoning`（自由記述）の 3 項目で構成される構造化応答を要求した。温度は 0.3、最大トークンはそれぞれ 256/512 に固定した。Reflection では所見抽出・鑑別診断・最終評価の 3 段階を書き分けるよう追加指示を与え、LLM 自身に思考の分節化を促した。Temporal Agent は過去の診療経過を圧縮した `state_summary`（300 字以内）、当日に観測された `new_findings`、および上記のリスクスコアと判定を返し、前回ステートを次回入力へ連結することで自己更新させた。いずれの条件も `is_suspected = true` または `risk_score >= 50` なら「兆候あり」と判定し、陽性・陰性で同じ基準を用いて感度と特異度を算出した。

表 3　実験条件

	Local	Global	Temporal
Lv1 Rule	1a	1b	–
Lv2 Zero-shot	2a	2b	–
Lv3 Reflection	3a	3b	3c

3.6　実験条件

Context 管理戦略とプロンプト設計の組み合わせにより、表 3 に示す 7 条件を設定した。

Temporal 戦略は Reflection プロンプトとの組み合わせ（3c: Temporal Agent）のみを実装した。State は「これまでの経過要約（300 字以内）」「今回の新規所見」「リスクスコア」「兆候判定」「説明」から構成し、各チェックポイントで更新しながら推論を進めた。

4　実験

4.1　設定

LLM には Qwen2.5-72B-Instruct-AWQ[14] を使用し、NVIDIA A100 80GB 上で vLLM により推論を行った。評価指標として、各時点（Day -30, -21, -14, -7, -4）における累積検出率（Sensitivity）および累積誤検出率（1 - Specificity）を算出した。`is_suspected=true` または `risk_score >= 50` であれば「検出」とみなし、陽性群に対する検出割合を感度、1-（陰性群の検出割合）を特異度とした。

4.2　結果

表 4 に各条件における累積検出性能を示す。

表 4　各条件における累積検出性能（N=92）

条件	Sens.	Spec.	PPV	NPV
1a Rule Local	60.9%	60.9%	60.9%	60.9%
1b Rule Global	60.9%	60.9%	60.9%	60.9%
2a ZS Local	52.2%	78.3%	70.6%	62.1%
2b ZS Global	56.5%	73.9%	68.4%	63.0%
3a Refl Local	**67.4%**	71.7%	70.5%	68.8%
3b Refl Global	65.2%	71.7%	69.8%	67.3%
3c Temporal	41.3%	**82.6%**	70.4%	58.5%

図 2 に F1 スコアの時系列推移を、表 5（付録）にチェックポイント別の感度を示す。Day -14 を境に F1 スコアが上昇し、Reflection (Global) が 68%、Zero-shot (Global) が 67%で最高となった。Local 戦

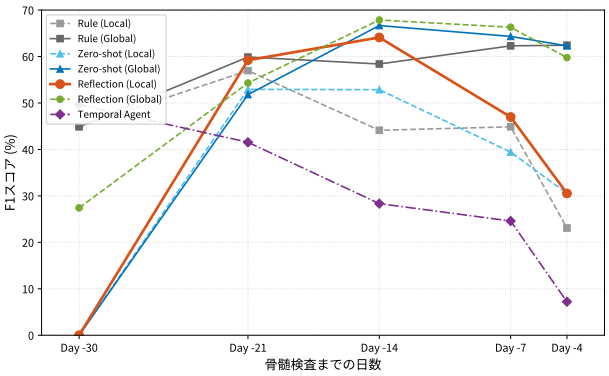

図 2　チェックポイント別 F1 スコアの推移。Day -14 で Reflection (Global) が 68%、Reflection (Local) が 64%を達成。Local 戦略は Day -14 前後がピーク、Global 戦略は検査直前まで高 F1 を維持。

略は Day -14 前後でピークを迎えた後に低下する一方、Global 戦略は検査直前まで高い F1 を維持した。Temporal Agent は時間経過とともに F1 が急降下し、状態圧縮の限界を示している。

RQ1: LLM による予測可能性　Reflection Local（3a）が感度 67.4%、特異度 71.7%、PPV 70.5%を達成し、Day -14 時点で陽性患者の 50%を検出した（表 5）。

RQ2: Context 管理戦略の比較　Local と Global は Reflection 条件で感度 67% vs 65%、F1 0.64 vs 0.68 と僅差だった。一方、Temporal Agent（3c）は感度 41%で、Day -7 時点の感度が 15%まで低下した。

RQ3: 偽陽性率　Rule-based（1a, 1b）は特異度 60.9%（偽陽性率 39.1%）となり、Zero-shot Local では偽陽性率 21.7%（特異度 78.3%）、Temporal Agent では 17.4%（特異度 82.6%）へ低減したが、感度は 41%にとどまった。

5　考察

5.1　早期検出の可能性と限界

チェックポイント別の分析から、Day -14（検査 2 週間前）が臨床的転換点であることが明らかになった。Reflection Local 条件ではこの時点で感度 50%・特異度 94%を達成し、骨髄像検査のオーダーを先行して検討できる実用的なシグナル強度を確保した。一方、Day -30〜-21 では感度 0–45%と低く、血液悪性腫瘍の臨床症状が骨髄検査の 1–2 週前に顕在化するという臨床知見と整合した。

症例分析から、検出成功は以下の 3 パターンに分類された：(1) 診断ワークアップパターン（生検や外注検査の結果記載がトリガー）、(2) 偶発的発見パターン（別疾患治療中に異常血算が発覚）、(3) 症状蓄積パターン（倦怠感・リンパ節腫脹・LDH 上昇など複数所見が閾値を超える）。LLM 出力は医師の診断プロセスと整合する傾向があり、確定診断結果が出る前の断片的な検査情報からの推論にも一定の成功を示した。

5.2　Temporal Agent の失敗分析

本研究の当初仮説は「時系列を追跡する Agent が最も高精度」であったが、結果は逆（感度 41%、特異度 83%）となった。出力分析の結果、陽性患者の 73%が「existing_followup」（既存疾患の経過観察）に分類されていた。これは、一度「既存フォロー」と判定されると、その後の悪化所見も軽視される「False Forgetting」問題による。

根本原因は 2 点ある：(1) **状態圧縮による情報消失**—Global 手法が Day -7 で感度 60%を達成したのに対し、Temporal Agent は 15%にとどまった。300 字への要約過程で診断的価値の高い情報が失われた可能性がある。(2)**「進行」vs「新規」の混同**—プロンプトで「既存疾患フォローなら疑いなし」と厳格に指示したことが、既存の良性血液疾患が悪性化したケースの見逃しにつながった可能性がある。

情報理論的な解析で示される自己修正の盲点 [13] と同様に、Agent の状態更新には外部検証機構が不可欠である。次世代設計として、構造化フラグによるハイブリッド State 管理や、ベースラインからの「逸脱検知」ロジックの導入を提案する。

6　おわりに

本研究では、骨髄像検査を受けた入院患者 92 名の匿名化カルテを基に、LLM による骨髄検査前兆候検出を評価した。Reflection Local 条件で感度 67%・特異度 72%を達成し、検査 2 週間前に陽性患者の半数を捕捉できる実効性を示した。一方、Temporal Agent は State 圧縮によって感度 41%に低下し、断面的な所見の組み合わせを重視する戦略が有効であることが確認できた。本稿で明示した時系列データ整備フローと再現可能な評価手順は、国内の臨床 NLP 研究における LLM 利用の基盤となる。

今後はマルチ施設検証、小規模モデルとの性能比較、Rule-based とのアンサンブル戦略を検討し、臨床導入に向けた信頼性向上を進める。

謝辞

本研究は東北大学病院 医療データ利活用センターおよび医療 AI センターの皆様によるデータ抽出・匿名化支援によって遂行された。ここに深謝する。

参考文献

[1] Parvati Naliyatthaliyazchayil, Raajitha Muthyala, Judy Wawira Gichoya, and Saptarshi Purkayastha. Evaluating the reasoning capabilities of large language models for medical coding and hospital readmission risk stratification: Zero-shot prompting approach. *Journal of Medical Internet Research*, Vol. 27, p. e74142, 2025.

[2] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-MedNLP: Overview of REAL document-based medical natural language processing task subtasks. In *Proceedings of NTCIR-16*, pp. 285–296, 2022.

[3] Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. JMedBench: A benchmark for evaluating japanese biomedical large language models. In *Proceedings of COLING 2025*, pp. 5918–5935, 2025.

[4] Gregory A. Abel, Christopher R. Friese, Lysa S. Magazu, Lisa C. Richardson, Maria E. Fernandez, Juan Jaime De Zengotita, and Craig C. Earle. Delays in referral and diagnosis for chronic hematologic malignancies: a literature review. *Leukemia & Lymphoma*, Vol. 49, No. 7, pp. 1352–1359, 2008.

[5] Constantinos Koshiaris, Jason Oke, Lucy Abel, Brian D. Nicholson, Karthik Ramasamy, and Ann Van den Bruel. Quantifying intervals to diagnosis in myeloma: a systematic review and meta-analysis. *BMJ Open*, Vol. 8, No. 6, p. e019758, 2018.

[6] Kazuya Shimoda, Naoto Takahashi, Keita Kirito, Noriyoshi Iriyama, Tatsuya Kawaguchi, and Masahiro Kizaki. JSH practical guidelines for hematological malignancies, 2018: I. leukemia-4 chronic myeloid leukemia (CML)/myeloproliferative neoplasms (MPN). *International Journal of Hematology*, Vol. 112, pp. 268–291, 2020.

[7] Kiyohiko Ohmachi. JSH practical guidelines for hematological malignancies, 2023: II. lymphoma-overview. *International Journal of Hematology*, Vol. 121, pp. 567–576, 2025.

[8] Maryam Gholipour, Reza Khajouei, Parastoo Amiri, Sadrieh Hajesmaeel Gohari, and Leila Ahmadian. Extracting cancer concepts from clinical notes using natural language processing: a systematic review. *BMC Bioinformatics*, Vol. 24, p. 405, 2023.

[9] Kenneth L. Kehl, Wenxin Xu, Eva Lepisto, Haitham Elmarakeby, Michael J. Hassett, Eliezer M. Van Allen, Bruce E. Johnson, and Deborah Schrag. Natural language processing to ascertain cancer outcomes from medical oncologist notes. *JCO Clinical Cancer Informatics*, Vol. 4, pp. 680–690, 2020.

[10] Gayathri Namasivayam, et al. Use of large language models to extract cancer diagnosis, histology, grade, and staging from unstructured electronic health records. *Journal of Clinical Oncology*, Vol. 43, p. 11176, 2025.

[11] Valentina D'souza, Alaleh Azhir, Danielle F. Pace, Samuel Friedman, Arash Nargesi, Tristan Naumann, Christopher D. Anderson, Steven J. Atlas, Judy Hung, and Mahnaz Maddah. CLIN-SUMM: Temporal summarization of longitudinal clinical notes. *medRxiv (preprint)*, 2025.

[12] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth M. Goldberg, and Yanjun Gao. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of EMNLP 2025*, pp. 20715–20735, 2025.

[13] Ken Tsui. Self-correction bench: Uncovering and addressing the self-correction blind spot in large language models, 2025.

[14] Qwen Team. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/, 2024.

A　チェックポイント別感度

表 5　チェックポイント別感度（Positive 群, n=46）

条件	Day-30	Day-21	Day-14	Day-7	Day-4
Rule Local	33%	45%	30%	35%	15%
Rule Global	33%	50%	47%	57%	59%
ZS Local	0%	36%	37%	28%	20%
ZS Global	0%	36%	53%	55%	57%
Refl Local	0%	45%	**50%**	35%	20%
Refl Global	17%	41%	57%	60%	52%
Temporal	33%	27%	17%	15%	4%
(n=)	6	22	30	40	46

Day -14（検査 2 週間前）を境に感度が上昇し、Reflection Local 条件では陽性患者の 50%を検出可能であった。

B　チェックポイント別特異度

表 6　チェックポイント別特異度（Negative 群, n=46）

条件	Day-30	Day-21	Day-14	Day-7	Day-4
Rule Local	86%	87%	94%	79%	85%
Rule Global	86%	83%	86%	74%	70%
ZS Local	86%	100%	97%	86%	89%
ZS Global	86%	97%	94%	84%	74%
Refl Local	93%	93%	**94%**	86%	89%
Refl Global	93%	90%	89%	79%	78%
Temporal	**100%**	**97%**	**97%**	**93%**	**93%**
(n=)	14	30	36	43	46

Temporal Agent は全チェックポイントで最も高い特異度を達成した。これは感度の低さ（表 5）と表裏の関係にあり、State 圧縮による保守的な判定が偽陽性抑制に寄与している。