MSR24

Predicting Treatment Response Depth in Multiple Myeloma Using Electronic Health Records

The Added Value of Large Language Model Extracted Clinical Features

Meenakshi DUBEY[1] • **Kok Joon CHONG**[1] • **Lee Yi FOO**[1] • **Yuba Raj PUN**[1] • **Kee Yuan NGIAM**[2,3]
Allison TSO[4] • **Hwee Lin WEE**[1] • **Melissa G. OOI**[3,5]

[1] Saw Swee Hock School of Public Health, NUS • [2] Department of Surgery, National University Hospital • [3] Yong Loo Lin School of Medicine, NUS
[4] Tan Tock Seng Hospital, Singapore • [5] National University Cancer Institute Singapore

Correspondence: Hwee Lin WEE • weehweelin@nus.edu.sg

354
patients in the cohort

0.76
best cross-validation AUC

0.72
best temporal-validation AUC

+0.14
absolute AUC gain, baseline to best

LLM-extracted clinical features, multivariate imputation and a clinically meaningful binary endpoint together produced the strongest discrimination.

Background

- Depth of response under International Myeloma Working Group (IMWG) criteria is an established surrogate for progression-free and overall survival, yet most myeloma prediction models target binary response only.
- Trial-standard predictors (FISH cytogenetics, R-ISS, ECOG, frailty, MRD) are often recorded only in free-text notes, so structured EHR data capture them incompletely.
- Response-depth prediction is largely untested in multi-ethnic Asian routine-care cohorts.

OBJECTIVE

To determine whether clinical features extracted from unstructured notes by a large language model (LLM) improve prediction of treatment response depth beyond structured EHR data alone.

Methods

- Retrospective cohort: 354 adults aged 21–100 years with confirmed multiple myeloma (ICD-9/-10, SNOMED) at National University Hospital, Singapore, January 2017 to December 2021.
- 34 variables from 6 enterprise data warehouse tables; conformance, consistency and plausibility checks passed for all variables.
- A locally trained Llama 3.1 extracted note-based variables; outputs were cross-checked against structured records.

MODEL A : ROUTINE EHR

Age, gender, ethnicity, MM subtype, β2-microglobulin, LDH, albumin, haemoglobin, sFLC ratio, ISS stage and medications.

MODEL B : MODEL A + LLM-EXTRACTED

Adds R-ISS, FISH cytogenetics (del(17p), t(4;14), 1q21 gain), ECOG performance status, frailty and CRAB symptoms from clinical notes.

MISSING-DATA STRATEGIES

Native handling by gradient boosting vs median–mode, 5-nearest-neighbour and multiple imputation by chained equations (MICE). MICE fills gaps by modelling each incomplete variable from all other variables over repeated cycles.

OUTCOMES AND VALIDATION

IMWG 5-class (CR, VGPR, PR, MR, PD), one-vs-rest, and binary deep response (CR+VGPR vs rest). HistGradientBoostingClassifier; 3-fold cross-validation and temporal validation (train 2017–2019, test 2020–2021).

Trial-standard vs routine-practice data

The main gaps are formal IMWG response adjudication, FISH panels at relapse, and consistent frailty documentation, assessments that are collected in clinical trial protocols..

INTENDED USE

Research-level applications: cohort construction, trial-eligibility screening and real-world health technology assessment.

Not a substitute for IMWG response assessment or for individual treatment decisions.

Results

STUDY COHORT · N = 354

64.7
median age, years (range 25–96)

55.1%
male

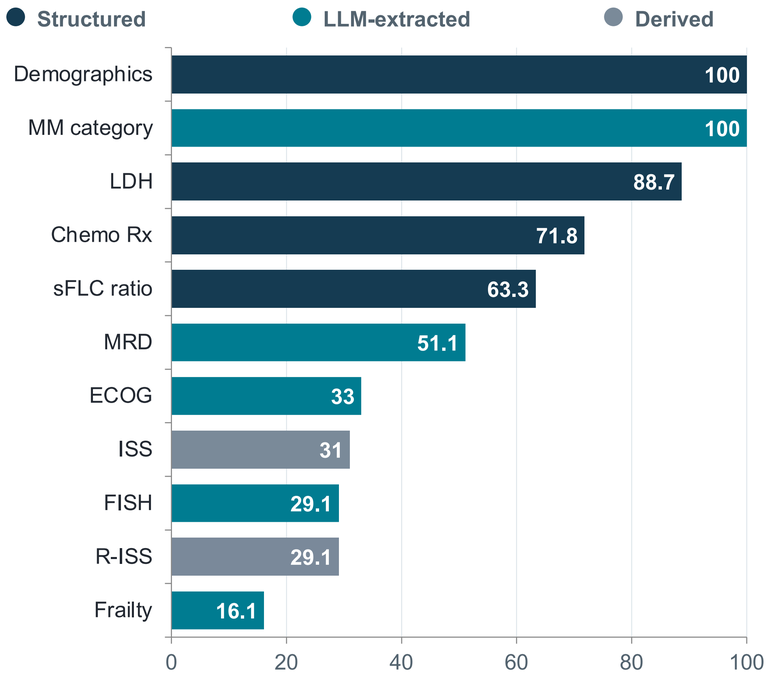
Ethnicity
Chinese 65.5%
Malay 16.7%
Indian 8.5%
Other 9.3%

Disease status at extraction
Newly diagnosed 49.4%
In remission 37.0%
Relapsed / refractory 13.6%

CRAB features at presentation
Bone disease 64.7%
Anaemia 57.3%
Renal impairment 26.0%
Hypercalcaemia 8.5%

Routine-care documentation was highly uneven

Completeness in full cohort (n = 354)



R-ISS and FISH are usually assessed only in active disease; among active-disease cases completeness was 45.7% and 41.7%.

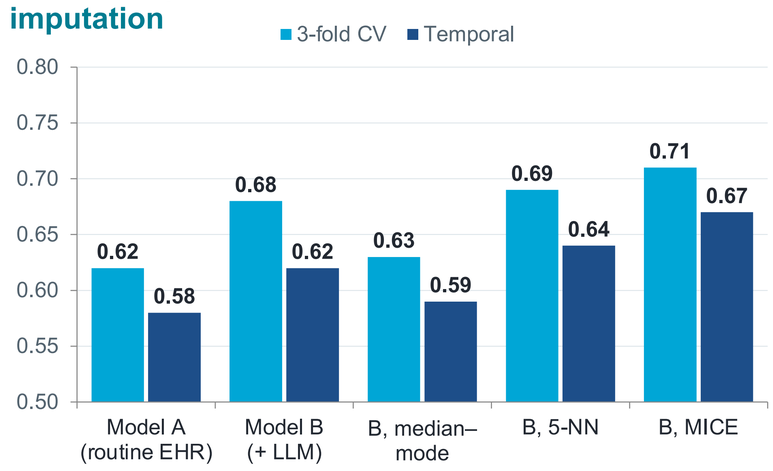
LLM extraction agreed with structured records

100%	96.3%	96.5%	92.9%
Medications	ISS staging	FISH	MRD status

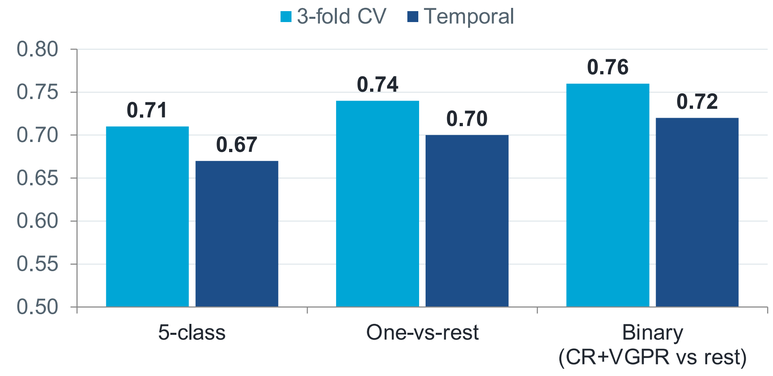
Concordance where both sources were available. The 7.1% discrepancy rate for MRD shows extraction is reliable but not error-free.

Model performance (macro-averaged AUC)

Five-class prediction: effect of LLM features and imputation



Outcome frameworks: Model B with MICE imputation



Baseline to best: +0.14 AUC in cross-validation (0.62 → 0.76) and temporal validation (0.58 → 0.72).

Feature set	5-class	One-vs-rest	Binary
Model A: routine EHR	0.62 / 0.58	0.64 / 0.62	0.64 / 0.64
Model B: + LLM features	0.68 / 0.62	0.69 / 0.67	0.70 / 0.69
Model B: + MICE	**0.71 / 0.67**	**0.74 / 0.70**	**0.76 / 0.72**

Values are AUC: 3-fold cross-validation / temporal validation. SD across folds ≤ 0.031 for all configurations; lowest for the binary model (0.009).

Key findings

- Adding LLM-extracted variables to routine EHR features raised 5-class AUC from 0.62 to 0.68 (cross-validation) and 0.58 to 0.62 (temporal validation).
- MICE imputation lifted performance further to 0.71 / 0.67, whereas naïve median–mode imputation degraded it below native missing-value handling (0.63 vs 0.68).
- The binary deep-response endpoint (CR+VGPR, a validated surrogate for progression-free survival) gave the best and most stable discrimination: AUC 0.76 / 0.72.
- Routine EHR variables alone reached binary AUC 0.64, so structured data carry useful baseline signal.

Limitations

- Single-centre retrospective design with a moderate sample size (n=354); limited power for rare response categories.
- High missingness constrains the effective sample size, and LLM extraction depends on documentation quality.
- Response labels came from clinical documentation rather than prospective adjudication; no external validation yet, and the 2017–2021 period predates newer regimens (D-VRd) and the IMWG 2025 criteria.

Future directions

- Multi-centre validation at Tan Tock Seng Hospital and Singapore General Hospital.
- Emulation of the PERSEUS trial in this real-world cohort; MRD as predictor and outcome under IMWG 2025 criteria.

CONCLUSIONS

Structured EHR data augmented with LLM-extracted clinical features distinguished deep response (CR+VGPR) from lesser responses with AUC 0.76 in cross-validation and 0.72 in temporal validation. This performance supports potential applications in trial-eligibility screening, treatment-selection research and real-world health technology assessment; however, external validation is required before patient-level use.

Selected references

1. Kumar S, et al. Lancet Oncol. 2016;17:e328–e346.
2. Palumbo A, et al. J Clin Oncol. 2015;33:2863–2869.
3. Mangal N, et al. Hematol Oncol. 2018;36:547–553.
4. Ferle M, et al. NPJ Digit Med. 2025;8:231.
5. Sonneveld P, et al. N Engl J Med. 2024;390:301–313.

Funding, disclosures and data

Funding: NMRC HPHSR Clinician Scientist Award (HCSASI21nov-0003).
Conflicts of interest: None declared.
Data availability: Protected health information; not publicly available.
Acknowledgements: NUHS Academics Informatics Office.

Contact

Hwee Lin WEE • weehweelin@nus.edu.sg
Saw Swee Hock School of Public Health,
National University of Singapore

AUC, area under the receiver operating characteristic curve; CR, complete response; VGPR, very good partial response; PR, partial response; MR, minimal response; PD, progressive disease; MICE, multiple imputation by chained equations; MRD, minimal residual disease; EHR, electronic health record; LLM, large language model; IMWG, International Myeloma Working Group; ECOG, Eastern Cooperative Oncology Group performance status; FISH, fluorescence in situ hybridisation; (R-)ISS, (Revised) International Staging System; sFLC, serum free light chain; LDH, lactate dehydrogenase; CRAB, hypercalcaemia, renal impairment, anaemia, bone disease; CV, cross-validation.