

Routine laboratory trajectories encode the onset of organ-level complications in cancer

Jannik Lübberstedt¹, Krischan Braitsch², Jacqueline Lammert^{3,4}, Christof Winter⁵, Florian Gabriel¹, Tristan Lemke¹, Christopher Zirn¹, Markus Graf¹, Friedrich Puttkammer^{1,6}, Hartmut Häntze⁶, Johannes Moll^{1,7}, Anirudh Narayanan^{1,6}, Andrei Zhukov⁶, Fabian Drexel¹, Zeineb Ben Chaaben¹, Sebastian Ziegelmayer¹, Su Hwan Kim¹, Marion Högner², Jan Kirschke⁸, Florian Bassermann^{2,9,10,11}, Marcus Makowski¹, Christian Wachinger¹, Lisa Adams^{1,*} & Keno Bressen^{1,7,*}

¹Department of Diagnostic and Interventional Radiology, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Germany.

²Department of Medicine III, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Germany.

³Chair of Medical Informatics, Institute of Artificial Intelligence in Medicine and Healthcare, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Germany.

⁴Clinical Department of Gynecology, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Munich, Germany.

⁵Department of Clinical Chemistry and Pathobiochemistry, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Germany.

⁶Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany.

⁷Department of Cardiovascular Radiology and Nuclear Medicine, School of Medicine and Health, TUM Klinikum, German Heart Center, Technical University of Munich.

⁸Department of Neuroradiology, School of Medicine and Health, TUM Klinikum, Rechts der Isar, Technical University of Munich, Germany.

⁹TranslaTUM, Center for Translational Cancer Research, Technical University of Munich, Munich, Germany.

¹⁰Deutsches Konsortium für Translationale Krebsforschung, Heidelberg, Germany.

¹¹Bavarian Cancer Research Center, Munich, Germany.

*Shared last authorship.

Abstract Routine laboratory panels drawn during cancer treatment constitute longitudinal physiological recordings of organ function, yet their temporal structure is discarded by single-timepoint prognostic tools. A transformer trained on 2,777,595 laboratory measurements from 3,905 patients with multiple myeloma or ovarian cancer predicted the two-year onset of 162 treatment-associated complications, including therapy-related myelodysplastic syndromes, spanning eight clinical categories, achieving 1.5- to 6.1-fold enrichment above prevalence at the group level. It matched or outperformed non-sequential baselines across grouped endpoints (AUROC gains up to +0.11), demonstrating that longitudinal laboratory trajectories capture evolving complication-specific physiology inaccessible from isolated measurements. Predictions generalised across both cancers, divergence concentrating in disease-specific complications, and biomarker masking recovered signatures consistent with established pathophysiology. External validation on MIMIC-IV and MMRF CoMMpass confirmed transferability across independent healthcare systems (AUROC up to 0.85). Routine oncological laboratory data encode organ deterioration weeks to months before clinical onset, enabling complication-specific surveillance without additional testing infrastructure.

Introduction

Patients with advanced cancer who receive prolonged multi-line therapy progressively accumulate organ-level complications across multiple systems [1], yet no existing model comprehensively predicts which specific complications a given patient will develop, or when [2, 3]. The problem is most acute in diseases that require years of sequential therapy in predominantly older populations, where cumulative toxicity compounds pre-existing organ vulnerability. Multiple myeloma and ovarian cancer are prototypical examples, as both impose prolonged multi-line treatment on patients with median ages at diagnosis of approximately 69 years in multiple myeloma and 64 years in ovarian cancer [4, 5], generating complication burdens that span renal dysfunction, marrow suppression, cardiovascular damage, recurrent infections, metabolic disease, and secondary malignancies [6, 7, 8, 9, 10].

Throughout this treatment course, routine laboratory panels are drawn at nearly every clinical encounter. Complete blood counts, metabolic panels, liver and kidney function tests, coagulation parameters, and inflammatory markers generate dense longitudinal records that reflect the state of multiple organ systems simultaneously. These measurements are universally available, impose no additional testing burden, and capture physiological changes that may precede clinical manifestation of complications. Yet this temporal signal remains underexploited.

The Cancer and Aging Research Group (CARG) score [2, 11] and the Chemotherapy Risk Assessment Scale for High-Age Patients (CRASH) [3] derive risk estimates from pre-treatment values; they neither update as physiology changes under therapy nor incorporate the trajectory of laboratory parameters across treatment lines, and they do not resolve individual complication types. External validation has reported discrimination as low as 0.52 (AUROC), in some populations [12, 13].

Disease-specific prognostic models, including recent transformer-based approaches for multiple myeloma that jointly predict survival and aggregate adverse events [14], do not resolve the onset of individual treatment-associated complications, leaving clinicians without actionable guidance on which specific toxicities to anticipate in a given patient. Generative transformers trained on large-scale electronic health records have demonstrated broad predictive capacity across general patient populations [15, 16, 17], but represent laboratory values as discretized tokens and require richly-coded event streams that vary substantially in completeness across institutions. Laboratory-trajectory models in cancer populations predict only a small number of aggregate endpoints [18]. No existing model applies raw laboratory trajectories to complication-resolved risk estimation within defined cancer populations. A recent systematic review identified only seven studies that applied machine learning to longitudinal clinical data for cancer complication prediction [19], confirming that the temporal dimension of patient monitoring remains largely unexploited by current modeling approaches [20, 21].

This study tested whether routine laboratory trajectories encode complication-specific temporal signal by training a transformer on 2,777,595 measurements from 3,905 patients with multiple myeloma or ovarian cancer, segmented into observation windows of 10–30 days, to predict the two-year onset of 162 treatment-associated complications. These two malignancies were chosen because they share the burden of cumulative treatment toxicity but diverge mechanistically in informative ways. Renal impairment is treatment-limiting in both yet arises from light-chain cast nephropathy in myeloma versus platinum nephrotoxicity in ovarian cancer [7, 22], and infection risk carries a disease-intrinsic component in myeloma, through clonal suppression of immunoglobulin production, that is absent in ovarian cancer [23]. This combination of shared and disease-specific endpoints provided a

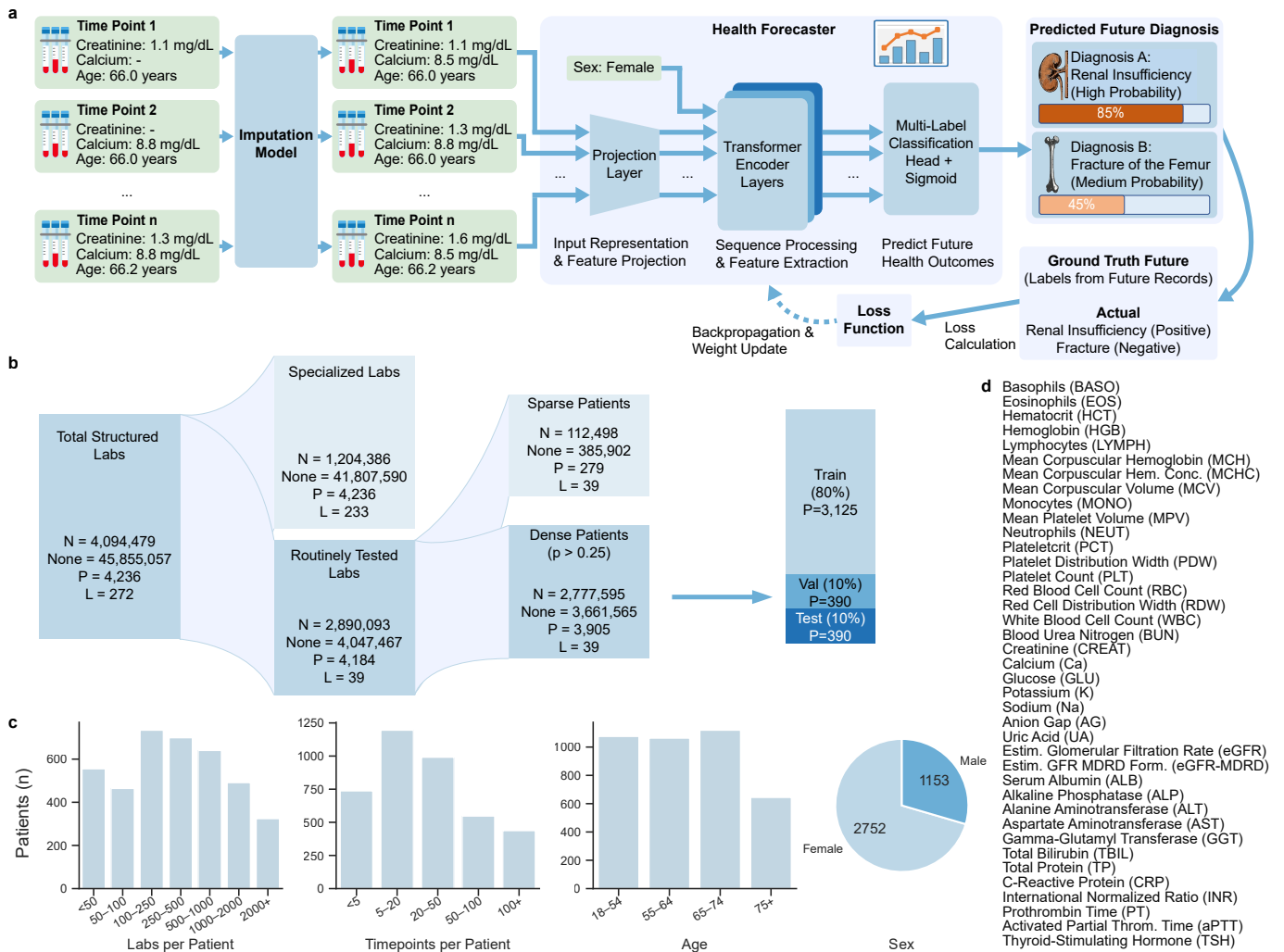


Figure 1. Transformer-based health forecasting model for diagnosis prediction from longitudinal laboratory data. **a**, Architecture of the diagnosis prediction pipeline. Longitudinal laboratory values are aggregated into tokens representing randomly sampled time intervals of 10–30 days, each encoding z-score-normalised values of the 39 routinely tested analytes listed in **d** together with patient age. Missing values are completed by a dedicated imputation model. Token representations are projected into the model's embedding space via a projection layer and processed by a transformer encoder. A multi-label classification head with sigmoid activation predicts 162 future diagnoses within 2 years after the input interval finished, trained against ground-truth labels derived from subsequent clinical records. **b**, Stepwise filtering of the laboratory dataset. Starting from all structured laboratory measurements ($N = 4,094,479$ values across $P = 4,236$ patients and $L = 272$ analytes), specialized laboratory tests were excluded to retain 39 routinely tested analytes ($N = 2,890,093$; $P = 4,184$ patients). Patients with sparse data coverage (proportion of observed values ≤ 0.25) were removed, yielding a final dataset of $N = 2,777,595$ measurements across $P = 3,905$ patients, split into training (80%, $P = 3,125$), validation (10%, $P = 390$) and test (10%, $P = 390$) sets. **c**, Cohort characteristics of the final dataset, showing distributions of the number of individual laboratory measurements per patient, number of aggregated time points per patient, patient age and sex. **d**, Complete list of the 39 routinely tested laboratory analytes used as model inputs, spanning haematology, metabolic, hepatic, inflammatory, coagulation and endocrine panel (Extended Data Table 2).

tested for whether laboratory-based prediction captures biologically coherent signal or relies on cohort-level confounders. Predictive performance was evaluated across both malignancies, compared against non-sequential baselines, interrogated through biomarker masking to identify which laboratory analytes drive complication-specific predictions, and tested for transferability against two independent United States cohorts, the MIMIC-IV critical-care database and the MMRF CoMMpass multiple myeloma registry. If this temporal signal exists, the data infrastructure for its detection is already in place in standard oncological care.

Results

Cohort characteristics

Across 39 routinely measured analytes, 2,777,595 structured laboratory measurements were extracted from the hospital information system of a single tertiary centre, spanning 3,905 patients (1,855 mul-

tiple myeloma, 2,050 ovarian cancer) treated between 2000 and 2026. Patients with sparse coverage ($<25\%$) were excluded and only parameters from standard clinical assessment retained (Figure 1b,c; Extended Data Table 2), then split into non-overlapping training ($P = 3,125$), validation ($P = 390$), and test ($P = 390$) sets. Clinical documents from the myeloma subcohort were separately processed by an LLM-based extraction pipeline for cohort characterisation only, not as model input (Supplementary Results, Extended Data Fig. 8; comparison against existing myeloma cohorts in Extended Data Table 1). Laboratory trajectories were segmented into non-overlapping 10 to 30 day intervals. Patients contributed a median of 16 timepoints (IQR 7–26) over 1,091 days (IQR 360–2,682). Missing values were completed by a transformer-based imputation model (see Methods); across the final dataset, 43.1% of analyte-timestep entries were observed. Diagnostic endpoints were curated from approximately 3,000 ICD-10 codes, retaining 162 diagnoses in eight categories: myelodysplastic syndrome, cardiovascular disease, kidney disease, fungal infections, bacterial

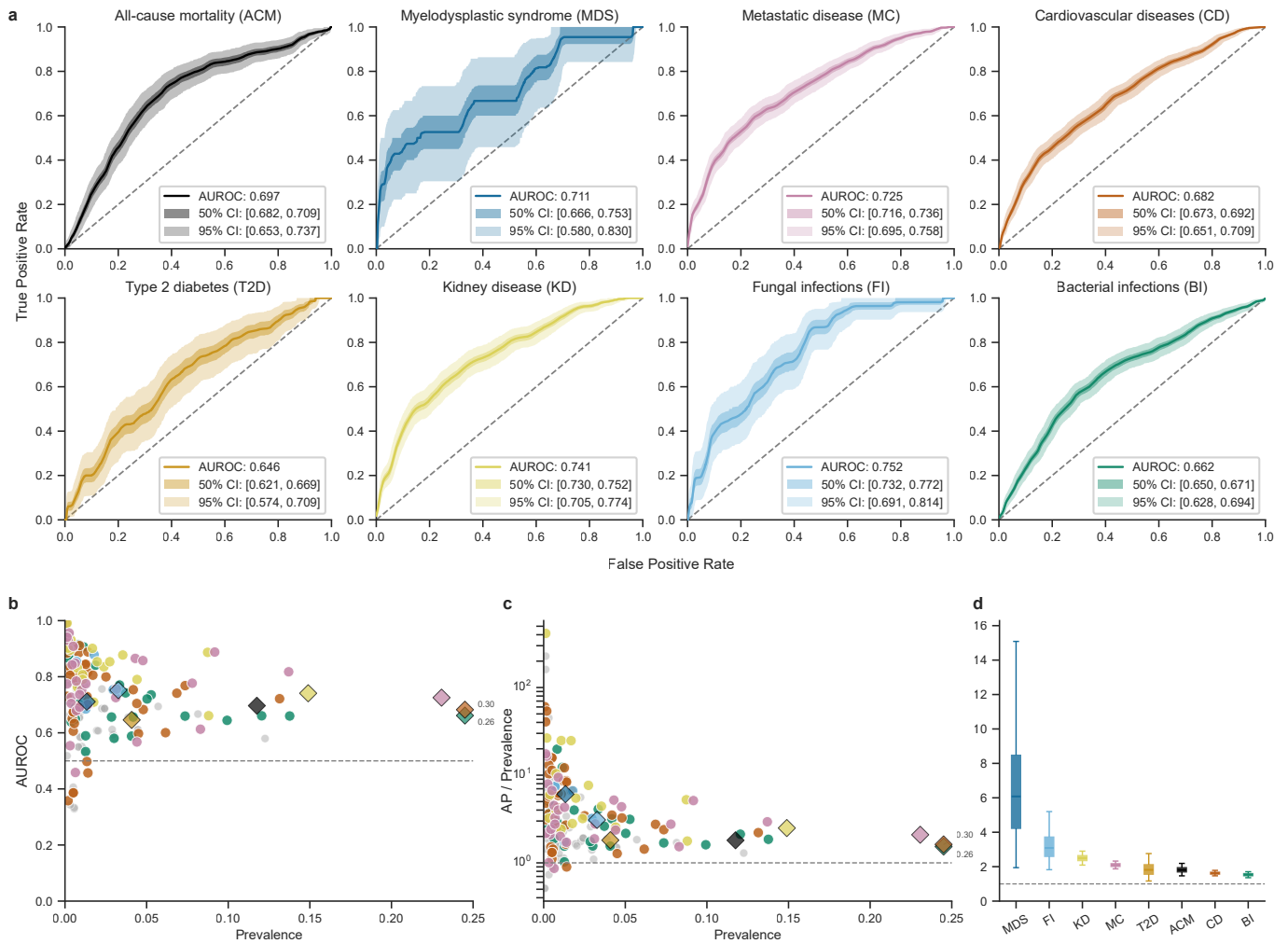


Figure 2. Diagnosis prediction performance across clinical endpoints and disease groups. **a**, Receiver operating characteristic curves for the prediction of eight clinically relevant diagnoses and disease groups within a 2-year horizon. Solid lines indicate median performance across bootstrap resamples. Shaded regions denote 50% and 95% bootstrap confidence intervals. The dashed diagonal represents random classifier performance (AUROC = 0.5). **b**, AUROC as a function of prevalence for individual ICD-10 diagnosis codes. Each point represents a single diagnosis; diamonds indicate grouped endpoint performance. Colours indicate disease group membership, grey shows without memberships in the eight primary groups. The dashed horizontal line at 0.5 indicates random discrimination. The cardiovascular disease (CD; prevalence = 0.30) and bacterial infection groups (BI; prevalence = 0.26) exceed the x-axis range and are plotted at the right edge with their prevalence indicated. Model discrimination consistently exceeds chance for higher-prevalence diagnoses, where estimation variance is lower. For rare diagnoses, confidence intervals widen but the model still achieves above-chance discrimination. **c**, Enrichment ratio (AP/prevalence, log₁₀ scale) as a function of prevalence for each individual diagnosis. The cardiovascular disease (CD; prevalence = 0.30) and bacterial infection groups (BI; prevalence = 0.26) exceed the x-axis range and are plotted at the right edge with their prevalence indicated. The model achieves the largest relative gains for rare diagnoses, with AP exceeding prevalence by up to an order of magnitude for the lowest-prevalence conditions. Diamonds indicate grouped endpoints. **d**, Enrichment ratio (AP/prevalence, linear scale) with 50% and 95% bootstrap confidence intervals for each grouped diagnostic endpoint, ordered by decreasing enrichment. The dashed line at 1.0 indicates the prevalence baseline. Endpoints above this line are predicted at above-chance enrichment, with myelodysplastic syndrome (MDS) and fungal infections (FI) showing the strongest relative gains.

infections, metastatic disease, type 2 diabetes, and all-cause mortality (Figure 1a). Performance on a held-out test set of 390 patients (1,579 sequences) was evaluated using AUROC and enrichment ratio (AP divided by prevalence) (Supplementary Tables 6–9; see Methods).

Prediction of treatment-associated complications

The transformer predicted the two-year onset of all eight primary disease groups above their respective prevalence baselines (Figure 2). AUROC ranged from 0.65 for type 2 diabetes to 0.75 for fungal infections, and predictive enrichment was inversely related to prevalence, with the highest ratios observed for the rarest endpoints, reaching up to an order of magnitude for diagnoses with prevalence below 0.01 (Figure 2b,c,d). The strongest enrichment among grouped endpoints was observed for myelodysplastic syndromes (AUROC 0.71; 6.1-fold, CI 95%=[1.9,15.1]), followed by fungal infections (AUROC 0.75; 3.1-fold), kidney disease (AUROC 0.74; 2.5-fold), and metastatic

disease (AUROC 0.73; 2.1-fold). All-cause mortality (AUROC 0.70; 1.8-fold), type 2 diabetes (AUROC 0.65; 1.8-fold), cardiovascular diseases (AUROC 0.68; 1.6-fold), and bacterial infections (AUROC 0.66; 1.5-fold) showed lower enrichment (Supplementary Table 7). Individual cardiovascular endpoints showed stronger enrichment and discrimination when evaluated separately, with chronic ischaemic heart disease reaching 2.5-fold enrichment (AUROC 0.71), atrial fibrillation 2.8-fold (AUROC 0.74), and heart failure 3.3-fold (AUROC 0.68).

Predictive performance across pre-diagnostic time horizons

To assess performance as a function of lead time, the model was evaluated cumulatively across forecasting horizons from 10 to 730 days (Figure 3). At each horizon, patients were labelled positive if the target diagnosis occurred between the end of the laboratory-input window and the specified prediction horizon.

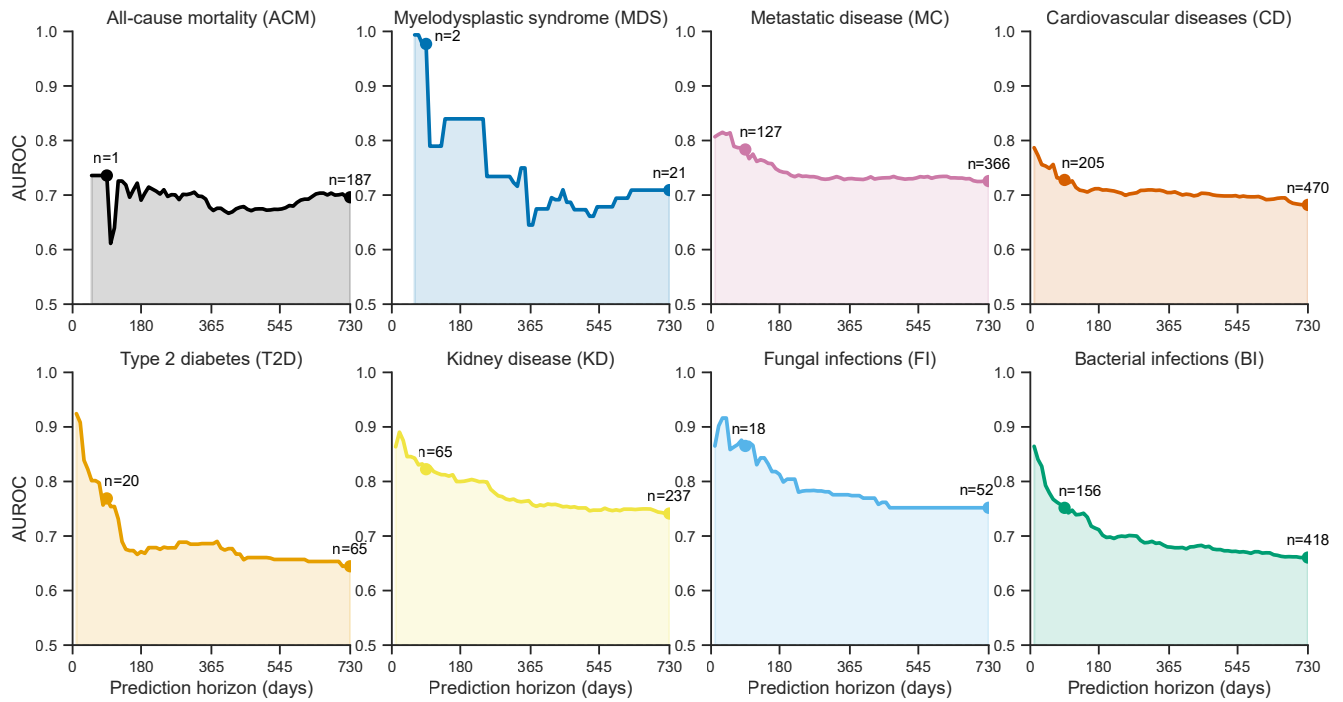


Figure 3. Cumulative diagnosis-prediction performance as a function of forecasting horizon. Area under the receiver operating characteristic curve (AUROC) for the Transformer-based health forecaster as a function of the prediction horizon, defined as the number of days after the end of the laboratory-input window within which the diagnosis occurs. At each horizon n , performance is evaluated cumulatively: a case is counted as positive if the diagnosis is recorded at any point from the end of the input window up to and including day n . The maximum prediction horizon is 730 days. Each panel corresponds to one of the eight grouped diagnoses. Curves show AUROC, with the filled region denoting the area under each curve. Annotations indicate the number of positive cases at 90 and 730s days.

Across the aggregate outcome, AUROC declined gradually with increasing forecasting horizon, from 0.682 at 10 days to 0.644 at 730 days. The grouped endpoints showed the same qualitative trend but differed in the rate of decline. Kidney disease, metastatic disease and fungal infections retained the highest long-horizon discrimination, remaining at AUROCs of 0.741, 0.725 and 0.752, respectively, at 730 days. Cardiovascular disease declined from 0.787 to 0.682 across the same interval, whereas bacterial infections and type 2 diabetes declined more steeply. All-cause mortality followed a distinct trajectory, becoming evaluable only after approximately 50 days owing to few early events, and stabilizing near 0.70 at longer horizons.

Biological coherence across malignancies

To test whether predictions reflected biologically coherent signal rather than cohort-level confounders, the model was evaluated separately on the multiple myeloma (MM; $P = 182$ patients, 853 sequences) and ovarian carcinoma (OC; $P = 208$ patients, 726 sequences) test subsets (Figure 4; Supplementary Table 7). Five of eight grouped endpoints generalised across both malignancies with comparable discrimination. Cardiovascular diseases (MM: AUROC 0.68; OC: AUROC 0.68), bacterial infections (MM: AUROC 0.64; OC: AUROC 0.68), metastatic disease (MM: AUROC 0.61; OC: AUROC 0.64), type 2 diabetes (MM: AUROC 0.56; OC: AUROC 0.63), and fungal infections (MM: AUROC 0.70; OC: AUROC 0.76) showed AUROC differences of no more than 0.07 across subcohorts despite differing prevalences. For metastatic disease, full-cohort discrimination (AUROC 0.73) exceeded both subcohort estimates, consistent with the model leveraging cohort-level prevalence differences in addition to within-cohort signal; C78 (secondary malignant neoplasms of respiratory and digestive organs) accounted for 28% prevalence in the ovarian subcohort but only 1% in myeloma, with correspondingly divergent average pre-

cision (OC: 0.40; MM: 0.02). Fungal infection discrimination was preserved across subcohorts despite a disease-asymmetric pathogen distribution, with pulmonary aspergillosis (B44.1) confined to the myeloma subcohort (MM: AUROC 0.84; OC: zero prevalence) and candidiasis (B37) captured in both populations (MM: AUROC 0.71; OC: AUROC 0.77). The remaining three endpoints diverged. Kidney disease showed a moderate cross-subcohort gap (MM: AUROC 0.69; OC: AUROC 0.79), while all-cause mortality showed a larger gap in the opposite direction (MM: AUROC 0.76; OC: AUROC 0.63). Myelodysplastic syndromes separated most sharply (OC: AUROC 0.93, 20.2-fold enrichment; MM: AUROC 0.50, 1.3-fold enrichment), with limited case counts ($N = 21$), precluding stable estimation.

Temporal advantage over non-sequential baselines

Two non-sequential baselines, XGBoost and logistic regression, were trained on tabular representations of the same laboratory sequences following the baseline evaluation protocol described in CoMET [17] (see Methods). The transformer matched or exceeded both XGBoost and logistic regression for all eight grouped endpoints, outperforming both baselines on six of them (Figure 5a; Supplementary Tables 6 and 7). The magnitude of the transformer's AUROC advantage over the best non-sequential baseline varied across endpoints. The largest gain was observed for myelodysplastic syndromes (transformer AUROC 0.71 vs. logistic regression 0.60; Δ AUROC +0.11), followed by fungal infections (AUROC 0.75 vs. XGBoost 0.66; +0.09). Cardiovascular diseases (AUROC 0.68 vs. logistic regression 0.63; +0.05), all-cause mortality (AUROC 0.70 vs. XGBoost and logistic regression, both 0.66; +0.04), and type 2 diabetes (AUROC 0.65 vs. logistic regression 0.61; +0.04) showed intermediate gains. Bacterial infections showed the smallest gain (AUROC 0.66 vs. logistic regression 0.65; +0.01). For metastatic disease (AUROC 0.73) and kidney disease

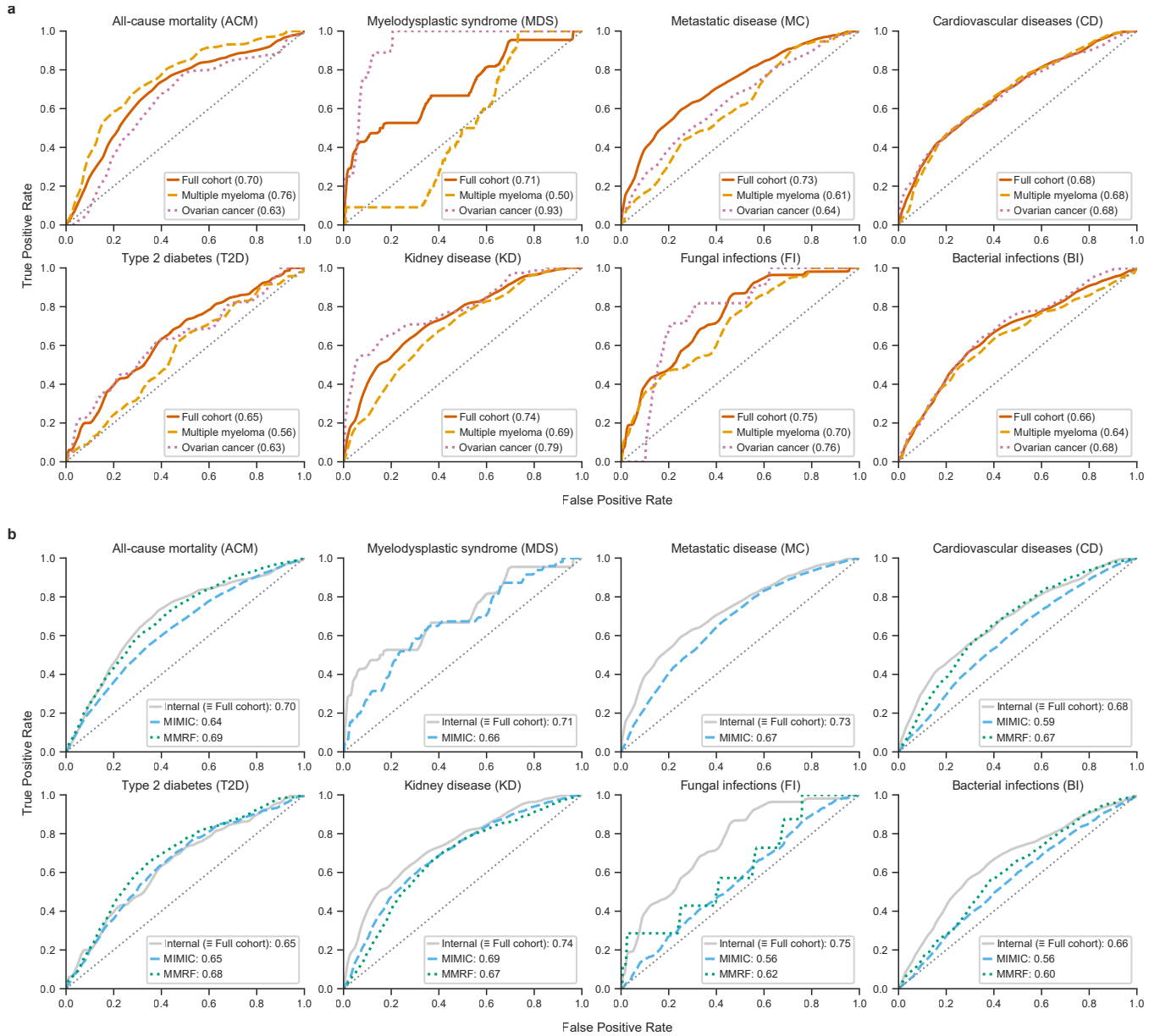


Figure 4. Diagnosis prediction performance across cancer-type subcohorts and Datasets. **a** Receiver operating characteristic (ROC) curves for the eight diagnosis groups evaluated on the full cohort (full cohort; $P = 390$ patients, 1,579 sequences), the multiple myeloma subcohort ($P = 182$, 853 sequences) and the ovarian carcinoma subcohort ($P = 208$, 726 sequences). Values in parentheses denote AUROC; the dashed diagonal represents random performance (AUROC 0.5). Five of eight endpoints generalised across subcohorts (AUROC differences ≤ 0.07); kidney disease, all-cause mortality, and myelodysplastic syndromes diverged. **b** ROC curves for eight diagnosis groups evaluated on the internal held-out test set ($N_{\text{eval}} = 1,579$ sequences), MIMIC-IV ($N_{\text{eval}} = 8,150$ sequences; v3.1) and MMRF CoMMpass ($N_{\text{eval}} = 7,347$ sequences; IA24). Values in parentheses denote area under the receiver operating characteristic curve (AUROC). The dashed diagonal represents random classifier performance (AUROC = 0.5). Discrimination is preserved across cohorts for all-cause mortality, type 2 diabetes and kidney disease, and additionally on MMRF for cardiovascular diseases, and on MIMIC-IV for MDS. Selective divergence on MIMIC-IV is observed for categories sensitive to coding-system differences, most prominently cardiovascular, bacterial and fungal endpoints. Myelodysplastic syndrome and metastatic disease are not shown for the MMRF subcohort, as both are not provided diagnostic labels.

(AUROC 0.74), the transformer performed on par with logistic regression (0.73 and 0.74, respectively). Model calibration is reported in Supplementary Results.

Biomarker signatures

To identify which laboratory analytes drove complication-specific predictions, single-feature and pairwise masking analyses were performed on the test set, recording the change in AUROC (Δ AUROC) when an analyte (or pair) and its validity indicator were zeroed across all timesteps (Figure 6).

Single-feature masking produced diagnosis-specific patterns that varied in concentration. Three endpoints had a clearly dominant an-

alyte: glucose for type 2 diabetes (-0.036), red blood cell count for myelodysplastic syndromes (-0.046 , the largest single drop in the matrix, ahead of the more conventional haemoglobin at -0.026), and the glomerular filtration estimates for kidney disease (eGFR -0.043 , eGFR-MDRD -0.039). For the remaining endpoints, discrimination loss was distributed across tight clusters of contributors rather than a single dominant biomarker (all-cause mortality, cardiovascular, metastatic, and fungal infections), with bacterial infections showing the weakest single-feature signal, led by eGFR. Across endpoints, albumin was the most cross-cutting contributor (top five in five of eight), eGFR led three (kidney, fungal, bacterial), and age led two (cardiovascular, all-cause mortality).

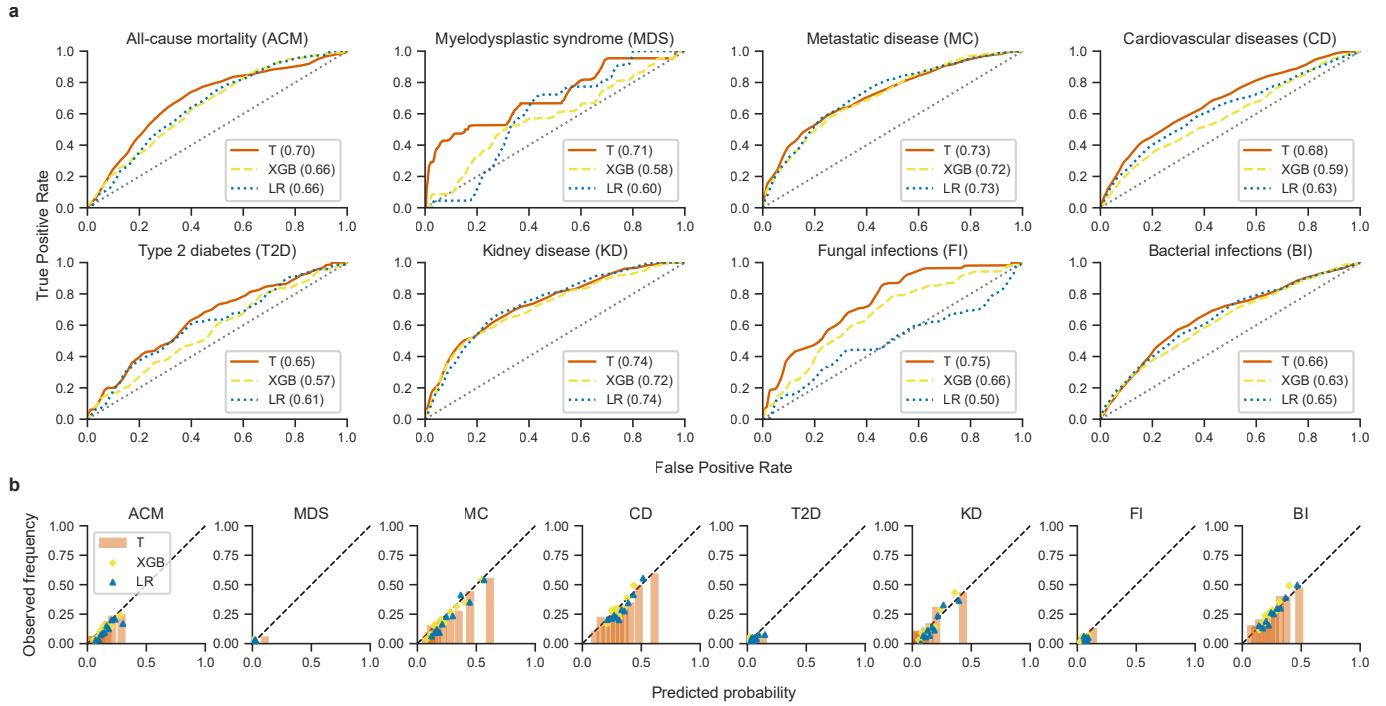


Figure 5. Model performance and calibration. **a**, Receiver operating characteristic curves for the transformer (T), XGBoost (XGB) and logistic regression (LR) baselines across eight diagnosis groups. Values in parentheses denote the area under the receiver operating characteristic curve (AUROC). The XGBoost and logistic regression architectures follow the implementations described in the *CoMET* framework, where XGBoost performed comparably to CoMET-L for incident disease prediction in low-prevalence settings. The transformer matches or exceeds both baselines on all eight endpoints, with gains over the strongest baseline of up to +0.11 for myelodysplastic syndrome and +0.09 for fungal infections; for metastatic and kidney disease it performs on par with logistic regression (AUROC 0.73 and 0.74, respectively). **b**, Per-endpoint calibration plots for all three models after Platt scaling on the validation set, fitted independently for each diagnosis group. Predictions are binned into ten equally sized quantiles. Each marker represents the mean predicted probability (x-axis) versus the observed event frequency (y-axis) within one quantile bin for the transformer (bars), XGBoost (diamonds) and logistic regression (triangles). The dashed diagonal indicates perfect calibration. All three models are well calibrated and overlap closely, tracking the diagonal across the observed range of predicted probabilities. For the rare endpoints (myelodysplastic syndrome, type 2 diabetes and fungal infections) predictions concentrate near the origin, consistent with low event prevalence, whereas for the more prevalent endpoints (all-cause mortality, metastatic disease, cardiovascular disease, kidney disease and bacterial infections) predictions span a wider range and remain close to the diagonal, with the highest-probability bins falling slightly below it.

Pairwise masking confirmed these patterns. The eGFR-MDRD/eGFR pair (−0.086) was the largest pairwise drop in the matrix, and all five top kidney pairs included a filtration estimate; MDS pairs all combined red-cell count with another marker (HGB/RBC −0.079). The strongest type 2 diabetes pair, glucose and age (−0.066), exceeded the sum of its singletons by 0.017, the largest super-additive excess of any top-ranked pair. Top pairs for the remaining endpoints combined the dominant singletons: age for cardiovascular (all five top pairs), eGFR for fungal and bacterial (four of five each), and INR for metastatic disease (all five, led by creatinine/INR −0.041).

Masking reflects both the predictive value of the biology and the model’s learned reliance on more frequently measured or correlated markers, and should not be read as intrinsic biomarker importance independent of training.

Robustness across subgroups

On the internal test set, removing sex, age, both, or laboratory imputation each changed the median AUROC by less than 0.005 from the full model (0.740), whereas ensembling contributed more substantially (−0.037); on the external MIMIC and MMRF cohorts the same ablations had larger effects and overall discrimination was lower (Figure 7). Subgroup analyses across age and sex are reported in the Supplementary Results.

External validation across independent healthcare systems

External validation was performed on the MIMIC-IV v3.1 database [24] and the MMRF CoMMpass IA24 cohort [25] under

out-of-distribution deployment without fine-tuning or cohort-specific recalibration. ICD-10-GM training labels were harmonized to ICD-10-CM and CoMMpass concept-level annotations using a clinically curated crosswalk (Supplementary Methods, Supplementary Table 3). Inference was performed across 8,150 MIMIC-IV and 7,347 CoMMpass sequences.

At the group level (Figure 4b, Table 9), transferability was strongest for renal and metabolic endpoints across both cohorts. Kidney disease transferred comparably across MIMIC-IV and CoMMpass (AUROC 0.694 and 0.668, respectively). Individual chronic kidney disease stages, end-stage renal disease and dialysis dependence achieved AUROCs up to 0.85 in MIMIC-IV, while renal and diabetic complication codes reached 0.74–0.76 in CoMMpass. Type 2 diabetes transferred comparably across cohorts (0.650–0.682). MIMIC-IV additionally retained discrimination for metastatic disease (0.667) and myelodysplastic syndrome (0.663), whereas these endpoints were not directly evaluable in CoMMpass because metastatic deposits and MDS are not explicitly represented in its coding structure.

Performance for all-cause mortality, cardiovascular disease, and infectious complications remained moderate across cohorts, while renal and metabolic endpoints showed the strongest transferability. Overall, external validation demonstrated that routine laboratory trajectories capture complication-specific physiological signatures that generalise across independent healthcare systems despite differences in patient populations and coding structure.

Discussion

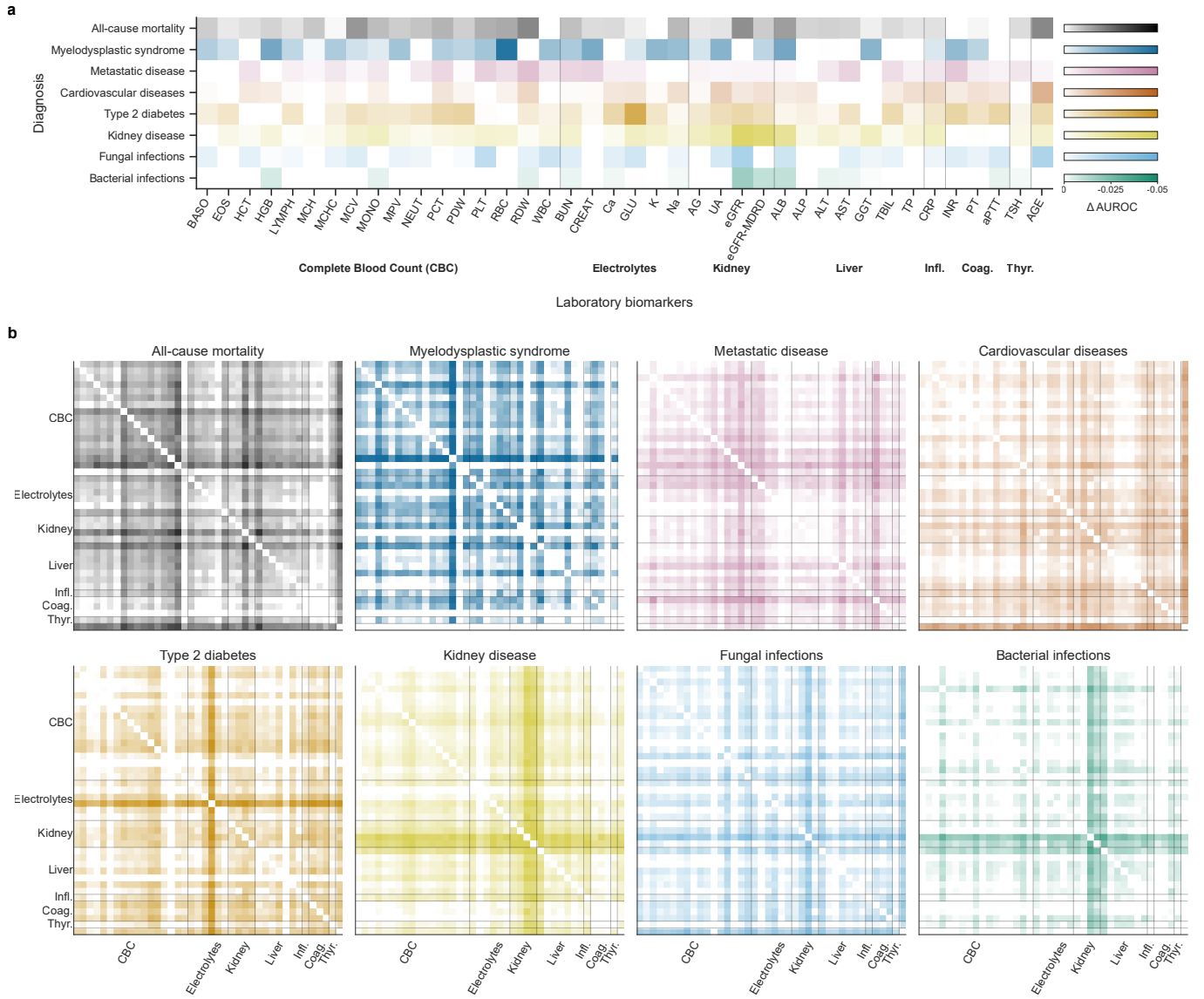


Figure 6. Feature importance analysis reveals diagnosis-specific biomarker contributions. **a**, Change in area under the receiver operating characteristic curve (Δ AUROC) when individual laboratory biomarkers or age are masked, shown per diagnosis group. Colour intensity reflects the magnitude of discrimination loss upon removal. Stronger shading indicates greater importance of the corresponding feature for predicting the given diagnosis. Features are grouped by laboratory panel (complete blood count, electrolytes, kidney, liver, inflammatory, coagulation and thyroid markers). **b**, Pairwise feature masking analysis. Each matrix shows the Δ AUROC when pairs of input features are simultaneously removed, for each of the eight predicted diagnosis groups. Off-diagonal entries reveal synergistic or redundant relationships between biomarkers. Strongly coloured off-diagonal cells indicate feature pairs whose joint removal causes a disproportionate discrimination loss beyond their individual contributions. Diagnosis-specific patterns emerge, such as the pronounced importance of kidney function markers for kidney disease prediction and of haematological parameters for myelodysplastic syndrome.

Short-term routine laboratory trajectories encode organ-level deterioration that precedes the clinical onset of treatment-associated complications. Predictive signal persisted across extended forecasting horizons for kidney disease, metastatic disease and fungal infections, despite gradual attenuation with increasing temporal distance from diagnosis. The advantage of temporal over cross-sectional modeling was most apparent for endpoints characterized by progressively evolving laboratory abnormalities rather than isolated measurements. The model predicted all eight endpoint groups above prevalence, with discrimination preserved (AUROC 0.71–0.75) for rare endpoints such as MDS and fungal infections where class imbalance typically degrades model performance. The predictions were biologically coherent across two mechanistically distinct cancers, temporally structured beyond what single-timepoint models could capture for most endpoints, and driven by interpretable biomarker signatures aligned with known pathophysiology.

Cardiovascular disease, bacterial infections, and fungal infections generalised across both malignancies with closely comparable discrimination, despite arising through different mechanisms in each cancer. Infection risk in myeloma carries a substantial disease-intrinsic component from humoral immunodeficiency, whereas in ovarian cancer it is predominantly treatment-mediated. The shared predictability reflects not shared biology, but shared laboratory observability. Where discrimination diverged, the gap tracks this same property, not tumour type.

All-cause mortality discriminated better in myeloma than in ovarian cancer (AUROC 0.76 vs 0.63) with the cause-of-death distribution accounting for most of the gap. Myeloma mortality is dominated by infection (immunoparesis [26], treatment-related neutropenia [27]), renal failure, and progressive cytopenias [28, 29], all of which manifest in routine panels. Ovarian cancer mortality is dominated by anatomic complications of peritoneal disease such as bowel obstruction and

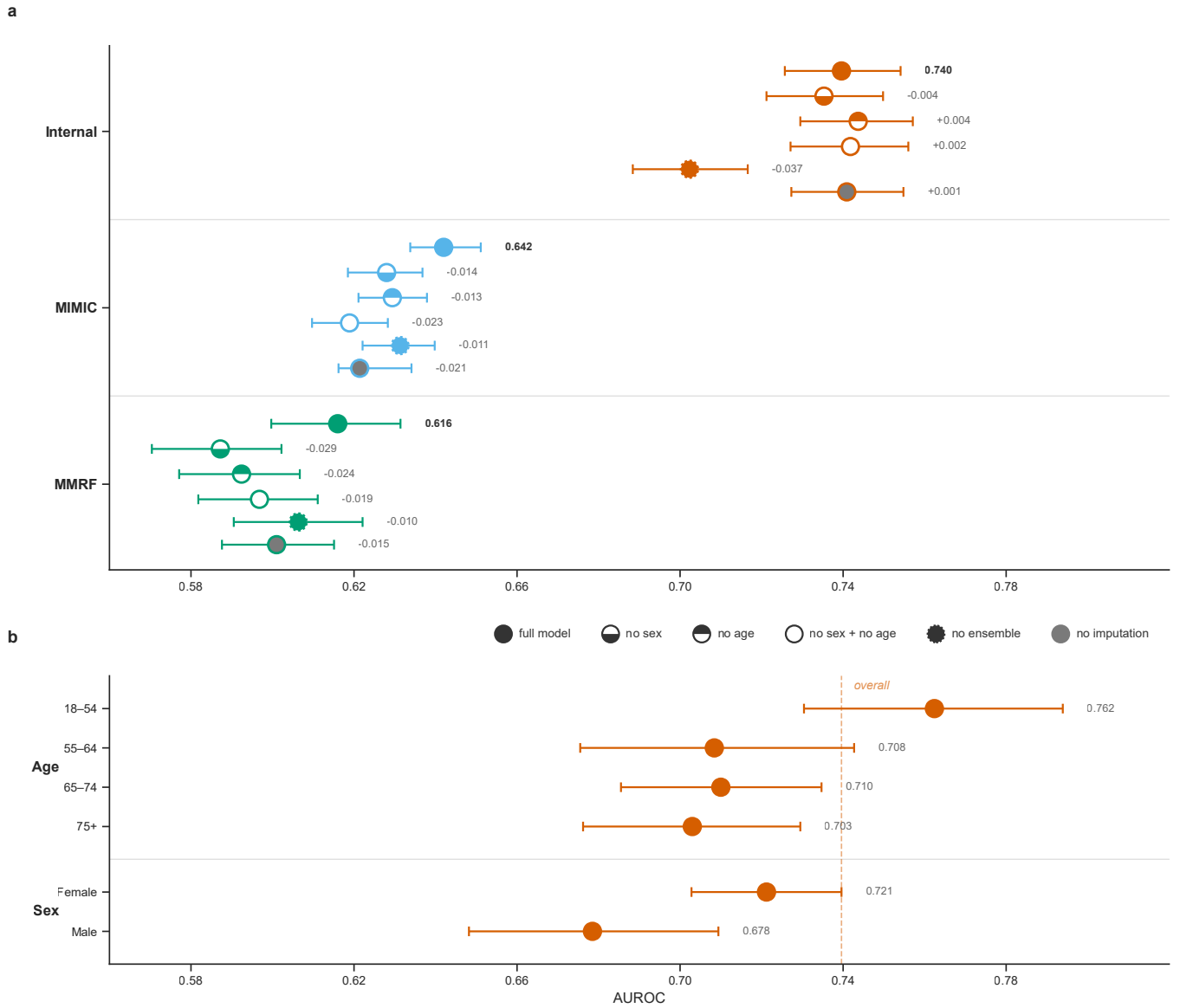


Figure 7. Subgroup and ablation analyses. Mean area under the receiver operating characteristic curve (AUROC) with 95% bootstrap confidence intervals across subgroups and ablations. For each ICD code, the median over 1000 bootstrap runs was computed; dots show the mean of these per-code medians, restricted to endpoints with ≥ 10 positives per bootstrap. **a** Performance for the full model and for retraining with sex, age, or both set to zero (value and validity marker). Gray dots indicate retraining without the imputation module, and dotted outlines retraining without model ensembling. Differences are minor on the internal dataset, except for ensembling (-0.037), but substantial on both external datasets (MIMIC, MMRF), indicating that demographics and the imputation module improve generalization. **b** Performance on the internal dataset stratified by age group and sex. The youngest stratum (18–54) shows the highest AUROC (0.762); the remaining age groups are comparable at a lower level (0.703–0.708). Discrimination is higher for female (0.721) than male (0.678) patients ($\Delta = 0.043$), yet both fall below the overall AUROC (0.740, dashed line), indicating that the full model's discrimination partly relies on sex differences.

cachexia [30, 31, 32], adjudicated by imaging and CA-125 rather than haematology or chemistry. The terminal decline is therefore more laboratory-visible in myeloma than in ovarian cancer.

The metastatic endpoints provided a harder test of the principle as of their different biology. Prevalences of lymph node (C77), respiratory and digestive (C78), and other-site (C79) secondary malignancies differed by an order of magnitude between cohorts, with C78 reaching 27.5% in ovarian cancer versus 1.2% in myeloma, reflecting that ovarian cancer spreads by solid-tumour routes [33] whereas myeloma does not metastasise in the same sense. Group-level discrimination was nevertheless near-identical for C78 (AUROC 0.60 vs 0.61) and C79 (0.60 vs 0.61), indicating that the model captures the laboratory signature of advanced disease burden in both cancers despite the biological difference between them. C77 diverged sharply (myeloma 0.83 vs ovarian 0.60), and the divergence again tracks observability. In

myeloma, C77 plausibly captures nodal extramedullary involvement with characteristic cytopenias [34]. In ovarian cancer, C77 captures heterogeneous nodal disease without a distinctive laboratory profile.

Myelodysplastic syndromes diverged sharply in the internal data (OC AUROC 0.93, MM AUROC 0.50), but 21 combined positive cases leave both estimates unstable. MIMIC-IV, with 41 MM cases, returned AUROC 0.66 and indicates a real signal in myeloma that the smaller internal sample missed. The underlying signature, pre-MDS cytopenias, macrocytosis, and elevated red-cell distribution width [35] arising from clonal haematopoiesis and therapy-related marrow injury [36, 37], is laboratory-observable by definition. However, if platinum [8] and PARP inhibitor [9] exposure amplify the signal in ovarian cancer relative to therapy-related marrow injury in myeloma cannot be resolved at this sample size.

The transformer's gain over non-sequential baselines varied from

+0.11 for myelodysplastic syndromes to zero for metastatic and kidney disease. Because the baselines and the transformer differ only in their capacity to model dependencies across timepoints, the variation tracks how much predictive information each endpoint carries in its trajectory rather than in the static distribution of laboratory values. The endpoints where the transformer gained most are slow-evolving. Myelodysplastic syndromes manifest as progressive cytopenias rather than a single abnormal value. Type 2 diabetes is preceded by glycaemic drift over an extended pre-diagnostic window [38]. For endpoints without a gain, the analysis cannot distinguish a weak temporal signal from one the non-sequential baseline captures equally well.

Together, the baseline and horizon analyses define which complications carry sufficient temporal signal for early risk stratification. For slow-evolving endpoints, the lead time is long enough that preventive intervention has room to act [39, 38].

Biomarker masking recovered diagnosis-specific signatures consistent with known pathophysiology, despite the model receiving no biological priors. Where a canonical laboratory marker exists, it dominated, namely glucose for type 2 diabetes and the glomerular filtration estimates for kidney disease, whose joint removal produced the largest pairwise effect of any endpoint. Because both estimates derive from creatinine and age, the single-feature drops are lower bounds on the contribution of filtration information. For myelodysplastic syndrome the signal was carried by anaemia (red-cell count and haemoglobin), with red-cell distribution width co-leading the broadly distributed mortality signature in line with its established prognostic association [40]. A small set of systemic markers, principally albumin and eGFR, recurred across endpoints, consistent with the model reading shared disease burden alongside endpoint-specific signal, and the strongest analyte pairs exceeded additive expectation, indicating that the model captures trajectory interactions rather than marginal effects. Notably, neutrophil count produced no measurable effect for bacterial infections, which were instead led by markers of chronic systemic deterioration and disease burden (eGFR, albumin, haemoglobin, and red-cell distribution width), consistent with a 730-day horizon targeting cumulative organ dysfunction and immune compromise making the patient more susceptible to infections, rather than acute neutropenic risk [23, 41].

Age dominated single-feature masking for cardiovascular disease but barely affected ablation. Retraining without sex or age produced negligible change internally (-0.004 , $+0.004$) and small drops on MIMIC (-0.014 , -0.013), with larger degradation on MMRF (-0.029 , -0.024). MMRF lacks eGFR and eGFR-MDRD, removing the reconstruction path through derived filtration markers. The reliance surfaced by masking thus reflects largely learned redundancy rather than unique information content, recoverable wherever the derived markers are present. Because these drivers are traceable to specific laboratory inputs, model behaviour supports clinical interpretability.

Complication-specific models for nephrotoxicity [42], febrile neutropenia [43], and cardiotoxicity [44] achieve strong discrimination individually but predict single endpoints from cross-sectional data and cannot be composed into a multi-organ profile. General-purpose EHR transformers such as Foresight [15], Delphi-2M [16], and CoMET [17] predict thousands of diagnoses but operate on coded events and tokenise laboratory values, requiring richly-coded event streams that most oncology departments do not routinely maintain. The closest comparator is SPARC [18], a laboratory-trajectory transformer predicting six endpoints in pan-cancer populations. The present model extends this direction to 162 diagnosis-specific endpoints within defined disease populations. For myelodysplastic syndromes, existing risk models such as CHRS [45] and MN-predict [46] achieve strong

discrimination but require next-generation sequencing, placing them outside routine clinical monitoring. Both nonetheless found routine CBC parameters carried independent predictive value [45, 46], consistent with the masking analysis reported here.

Predictive enrichment for myelodysplastic syndromes could identify high-risk patients for intensified haematological surveillance or emerging prevention strategies, such as CDK4/6 inhibition to limit chemotherapy-induced expansion of TP53-mutant clonal haematopoiesis [47], in a setting where monitoring currently relies on periodic blood counts without systematic risk stratification [9]. This requires no testing beyond standard oncological care. However, enrichment over prevalence demonstrates that predictive signal exists but does not establish the discrimination needed for individual management; translating it into actionable risk scores will require prospective calibration studies defining decision thresholds for specific interventions.

Cross-system external validation tested transfer to two structurally distinct cohorts without fine-tuning or recalibration. Type 2 diabetes transferred most stably, matching internal discrimination on MIMIC-IV (AUROC 0.65) and exceeding it on MMRF CoMMpass (0.68), while all-cause mortality and cardiovascular disease were largely retained on MMRF despite attenuation on MIMIC-IV. On MIMIC-IV, myelodysplastic syndrome and metastatic disease were best preserved, with MDS retaining 2.58-fold enrichment over prevalence. Infection endpoints degraded most on MIMIC-IV (bacterial 0.56, fungal 0.56) but recovered partially on MMRF (0.60, 0.62). Across both cohorts, all evaluable endpoints maintained enrichment at or above prevalence, indicating transferable laboratory signal even where discrimination was attenuated. The variable degradation reflected at least three structural coding differences between ICD-10-GM and the target systems, including obligatory versus optional secondary diagnosis coding for pathogens, non-equivalent cardiovascular and renal disease subcategorisation, and divergent reimbursement incentives that drive differential capture of comorbidities across DRG systems [48, 49]. Disentangling these effects from genuine cross-population shift is not possible with cross-sectional ICD coding alone, and the residual question of how much each contributes is a structural problem for every ICD-coded prediction model deployed across national systems.

This study has limitations. No prospective validation has been performed. Although the model was developed at a single center, the identified laboratory signatures transferred to two independent external cohorts (MIMIC-IV and MMRF CoMMpass), supporting generalizability beyond the development population. The model receives no treatment information, and the observed associations may therefore reflect disease progression, comorbidity burden, or treatment effect. Disentangling these contributions will require models that integrate treatment exposure data. The steeper decline in discrimination for bacterial infections across forecasting horizons suggests that their predictive signal is concentrated closer to clinical onset. Accordingly, the 730-day prediction horizon is appropriate for gradually evolving endpoints but may limit clinical actionability for acute infectious complications, where shorter horizons and more frequent model updates would constitute a distinct modeling task. The reliance on ICD-10-coded diagnoses as outcome labels introduces potential misclassifications, as demonstrated by the structural coding differences identified in the cross-system validation, and coding accuracy, completeness, and temporal assignment of diagnosis dates vary across institutions and over time. The 25-year observation window spans multiple changes in laboratory assays, treatment standards, and coding practices that may introduce confounders not captured by the current modeling framework. PARP inhibitors entered routine first-line use

in the final five to seven years of the observation period. Given the latency between exposure and therapy-related myelodysplasia, the MDS signal in the ovarian subcohort more plausibly reflects platinum toxicity. Characterising PARP-associated myelodysplasia [50] would require cohorts with longer, better-characterised exposure.

In conclusion, routine laboratory trajectories collected during standard cancer treatment encode complication-specific organ deterioration weeks to months before clinical onset. The signal sits inside data that every oncology unit already produces and every hospital information system already stores, making laboratory trajectory modeling the cheapest dense longitudinal monitoring channel available in oncology, with no new assay, sequencing run, or imaging study required to deploy it. The same trajectory is the natural substrate for patient-level digital twins of cancer, where a continuously updating organ-state representation is required and no competing modality delivers comparable temporal density at comparable cost. The 162-endpoint resolution shown here demonstrates that this substrate carries complication-specific rather than purely prognostic information, with interpretable biomarker signatures that keep the predictions auditable rather than opaque.

Methods

Study cohort and clinical data

A cohort of 2,328 patients with multiple myeloma treated at a tertiary care centre (Klinikum rechts der Isar, Technical University of Munich) between January 1st 2000 and January 1st 2026 was combined with 2,141 patients with ovarian cancer from the same institution. For both subcohorts, laboratory measurements and ICD-10-coded diagnoses used as input to the diagnosis prediction model were obtained from the hospital information system. Of these 4,469 patients, 4,236 had valid laboratory measurements recorded. After restricting analytes to 39 routinely tested parameters, 4,184 patients were retained; subsequent removal of patients with more than 75% missing values across retained analytes yielded the final cohort of 3,905 patients used for diagnosis prediction (Figure 1b). Independently, for the myeloma subcohort, a five-stage LLM-based information extraction pipeline using locally deployed Llama 3.3-70B [51] was applied to 90,036 unstructured German-language clinical documents, including 31,790 internal discharge summaries and specialised reports from radiology, pathology, and nuclear medicine departments. This pipeline produced a structured myeloma dataset encompassing treatment lines, cytogenetic assessments, bone marrow infiltration, refractory status, and clinical endpoints (Supplementary Methods, Extended Data Fig. 8, Supplementary Tables 4, 5). This dataset contextualises the cohort but was not used as model input. For the ovarian cancer subcohort, equivalent clinical metadata were available directly from the hospital information system.

Diagnosis prediction from laboratory trajectories

Dataset construction. The initial dataset comprised 4,094,479 laboratory measurements across 4,236 patients and 272 analytes. Input features were restricted to 39 analytes routinely obtained during standard clinical assessment, spanning haematology, metabolic, hepatic, coagulation, inflammatory, and endocrine panels (Extended Data Table 2), rather than tests ordered in response to specific clinical suspicion. This restriction reduces indication bias and limits predictions to measurements available in routine clinical care regardless of institutional infrastructure. After analyte filtering and removal of patients with more than 75% missing values, 2,777,595 measurements across 3,905 patients remained (Figure 1b).

Patient laboratory trajectories were segmented into non-overlapping temporal intervals with durations sampled uniformly between 10 and 30 days. Laboratory values obtained on the same calendar day were averaged per analyte to mitigate indication bias, as same-day repeat measurements typically reflect acute clinical concern rather than the gradual physiological trends targeted by the 730-day prediction horizon. Each day containing at least one valid measurement was represented as a single input token encoding the 39 analyte values together with the patient's age in days. A binary sex token was appended once per sequence. Intervals without measurements were discarded. The number of sequences per patient was capped at ten to limit overrepresentation of heavily monitored individuals, with the cap increased to 20 for the myelodysplastic syndrome endpoint to compensate for its low prevalence (Supplementary Methods). To prevent label leakage, sequences were excluded on a per-diagnosis basis if the target diagnosis had been documented prior to the first day following the end of the laboratory observation window.

Clinical outcomes were defined using 162 ICD-10 codes curated by a board-certified oncologist from approximately 3,000 distinct codes documented across the cohort. Each diagnosis was required to occur in at least 50 patients across the full dataset. Diagnosis prediction was formulated as a multi-label classification task estimating the probability of each diagnosis within 730 days following the end of the laboratory observation window. This horizon captures complications that manifest across one to two lines of therapy in both diseases [52, 9] (see Supplementary Methods for detailed justification). The final cohort was split at the patient level into training (3,125 patients; 12,145 sequences), validation (390 patients; 1,453 sequences), and test (390 patients; 1,579 sequences) sets.

Laboratory value imputation. Routine laboratory panels rarely measure all 39 analytes simultaneously, and the pattern of missingness may itself encode clinical context. A dedicated transformer-based imputation model was therefore trained prior to diagnosis prediction. The imputation model produced dense timestep representations, preventing the prediction model from exploiting differential test-ordering patterns rather than physiological values. The imputation model followed a masked prediction objective in which individual analytes were randomly masked with probability 0.2 and entire timesteps with probability 0.05. The model was trained to reconstruct masked values using mean squared error loss (architectural details and hyperparameters in Supplementary Table 10).

Prediction model. Temporal dependencies across laboratory timesteps were modelled using a bidirectional transformer encoder adopting the architectural components of Qwen-3 [53], including rotary positional embeddings, SwiGLU activations, RMSNorm, and grouped-query attention. These design choices reflect current best practices for capturing long-range dependencies in sequential data and have been validated extensively at scale in the language domain. The same inductive biases, particularly the capacity to model complex interactions across ordered sequences, were hypothesized to transfer to longitudinal laboratory trajectories. All weights were trained from scratch on structured laboratory data without language pretraining. Bidirectional attention was chosen because the prediction task operates on completed observation windows rather than generating sequential, token-wise predictions. Each laboratory interval is fully observed before diagnosis probabilities are computed, and bidirectional context allows the model to interpret early measurements in light of later values within the same interval. For each interval, normalised and imputed laboratory values were projected into a continuous embedding space

via a linear input layer. A linear prediction head mapped the attention-pooled final-layer representations of the transformer to multi-label diagnosis probabilities. Models were trained using weighted focal loss with class weights derived from training set diagnosis frequencies, and optimised with Adam using cosine annealing (architectural details and hyperparameters in Supplementary Table 11). The pipeline is illustrated in Figure 1a.

Evaluation. The combined training and validation sets were partitioned into five patient-level folds. One model was trained per fold, with four folds used for training and the held-out fold for validation. Hyperparameters were selected based on mean area under the receiver operator curve (AUROC) across validation folds. For final evaluation on the held-out test set, predictions were obtained by averaging raw logits across all five fold-specific models before applying the sigmoid function.

AUROC was chosen as the primary discrimination metric because it is threshold-independent, comparable across endpoints with differing prevalences, and widely reported in the literature, facilitating comparison with existing models. To contextualise discrimination relative to baseline event rates, predictive enrichment was quantified as the ratio of average precision to endpoint prevalence (enrichment ratio). Average precision is additionally reported for all endpoints in Supplementary Tables 6 and 7, as it directly reflects the clinical use case of identifying affected patients under class imbalance [54].

Group-level predictions were constructed by taking the maximum predicted probability across constituent diagnoses. Confidence intervals were obtained from 1000 nonparametric bootstrap resamples of the test set (Supplementary Methods).

Baselines. To isolate the effect of temporal modeling, the transformer was compared against two non-sequential baselines, logistic regression (linear) and XGBoost (non-linear), following the baseline methodology described in the CoMET benchmark [17]. Each laboratory sequence was converted into a bag-of-words count vector in which each feature is the number of measurements falling in a given (analyte, value-bin) cell, with values binned into deciles per analyte. Logistic regression features were scaled to unit maximum absolute value on the training set; XGBoost, being invariant to monotonic per-feature rescaling, consumed raw counts. Both baselines received the same demographic features (age and sex) and training labels as the transformer. XGBoost was included as the strongest supervised baseline, as it matched or exceeded CoMET-L on incident disease prediction tasks with low endpoint prevalence in that benchmark. This representation preserves each analyte’s binned value distribution over the window but discards the order and timing of measurements, so the transformer’s improvement over these baselines quantifies the predictive information carried by the temporal trajectory beyond an order-free summary of the same values.

External validation. Cross-system transferability was assessed on two structurally distinct external cohorts. First, MIMIC-IV (version 3.1), a publicly available critical care database from Beth Israel Deaconess Medical Center, Boston, USA. Patients were selected by ICD-10-CM diagnosis codes corresponding to multiple myeloma (C90.0x) or ovarian cancer (C56.x), yielding 1,626 patients (930 multiple myeloma, 697 ovarian cancer; one patient carried both diagnoses) comprising 8,150 sequences. Laboratory values for the same 39 analytes used in the primary analysis were extracted and mapped to their MIMIC-IV equivalents using manually curated crosswalks (Supplementary Table 2). Two of the 39 analytes (plateletcrit and

platelet distribution width) were unavailable in MIMIC-IV and were treated as missing inputs throughout that cohort. Diagnostic endpoints were mapped from the 162 ICD-10-GM targets to ICD-10-CM codes through manually curated crosswalks verified against the codes present in MIMIC-IV (Supplementary Table 3). No official crosswalk exists between ICD-10-GM and ICD-10-CM, as both are independent national modifications of WHO ICD-10 [48]; mappings were therefore curated by clinical review. Second, the MMRF CoMMpass IA24 release, a prospective longitudinal cohort of newly diagnosed multiple myeloma patients with serial laboratory monitoring, comprising 1,143 multiple myeloma patients yielding in 7,347 sequences. Laboratory values for the same 39 analytes were extracted and mapped to MMRF’s native laboratory concepts using manually curated crosswalks (Supplementary Table 2); unavailable analytes were treated as missing inputs throughout. Diagnostic endpoints were harmonised from the 162 ICD-10-GM targets to MMRF’s concept-level coding through manually curated mappings (Supplementary Table 3); metastatic codes and myelodysplastic syndrome had no equivalent concept, and were excluded from evaluation. For both cohorts, evaluation was performed on the entire dataset, and group-level and per-endpoint AUROC were computed using the same procedures as for the internal test set.

Biomarker importance analysis. Diagnosis-specific laboratory signatures were quantified through a leave-one-out masking approach. For each of the 39 analytes and age, and for each diagnostic group, the feature value and its corresponding validity indicator were set to zero across all timesteps in the test set, distinguishing masked entries from true zero measurements. The change in AUROC relative to the unmasked baseline was recorded. Pairwise masking was performed analogously for all $\binom{40}{2} = 780$ feature pairs to identify synergistic biomarker combinations.

Software and hardware. All models were implemented in Python using PyTorch 2.8.0 and Hugging Face Transformers 4.57.0. Evaluation metrics were computed using scikit-learn 1.7.2. Training was performed on a single Nvidia H100 GPU.

Data availability

The clinical dataset underlying this study contains protected health information and cannot be shared publicly under applicable data protection regulations (EU General Data Protection Regulation).

Code availability

Source code for the structuring, imputation and prediction model training and evaluation, together with trained model weights, will be made available at <https://jalu1870.github.io/lab-forecast/> upon publication.

Acknowledgements

This work is supported by the Bavarian State Ministry of Health, Care and Prevention through the GO-TWIN project (grant number DGP-2024-01).

Author contributions

J.L. conceived the study, planned and conducted all experiments for clinical data structuring and diagnosis prediction, constructed the information extraction pipeline, designed the model evaluation framework

and statistical analyses, performed input preprocessing, and wrote the manuscript. K.Bra. reviewed the manuscript, provided clinical expertise on multiple myeloma, contributed to endpoint definition for clinical data structuring, and selected diagnoses for prediction. J.La. reviewed the manuscript and provided clinical expertise on ovarian cancer. C.Wi. provided access to laboratory data. F.G. performed data annotations and advised on the design of the structuring study. T.L. and C.Z. performed data annotations. M.G. contributed expertise on multiple myeloma endpoint definition. F.P., A.Z., H.H., A.N., F.D., Z.B.C., J.M., S.H.K., S.Z., F.B., M.H., J.K., and M.M. critically reviewed the manuscript and provided feedback. C.Wa. supervised the project and recommended the XGBoost comparison analysis. L.A. and K.Bre. jointly supervised this work. They provided computational resources, advised on experimental design, figure design, and template construction, and contributed to manuscript editing and revision.

Competing interests

FB received honoraria from Amgen, Johnson & Johnson, Bristol Myers Squibb (BMS), AbbVie, and GSK, travel support from Amgen, Johnson & Johnson, and BMS, and served on advisory boards for Amgen, Johnson & Johnson, BMS, AbbVie, and GSK. MH received honoraria from Johnson & Johnson, Sanofi, GSK, Oncopeptides, and Pfizer, payment for expert testimony from Johnson & Johnson and Oncopeptides, travel support from Johnson & Johnson, Oncopeptides, Pfizer, and Amgen, and served on advisory boards for Johnson & Johnson, Sanofi, Oncopeptides, and Pfizer. All other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Jannik Lübberstedt (jannik.luebberstedt@tum.de).

References

- [1] Maryam B. Lustberg, Nicole M. Kuderer, Aakash Desai, Cristiane Bergerot, and Gary H. Lyman. Mitigating long-term and delayed adverse events associated with cancer treatment: implications for survivorship. *Nature Reviews Clinical Oncology*, 20(8):527–542, 2023.
- [2] Arti Hurria, Kayo Togawa, Supriya G. Mohile, Cynthia Owusu, Heidi D. Klepin, Cary P. Gross, Stuart M. Lichtman, Ajeet Gajra, Smita Bhatia, Vani Katheria, Shira Klapper, Kurt Hansen, Rupal Ramani, Mark Lachs, F. Lennie Wong, and William P. Tew. Predicting chemotherapy toxicity in older adults with cancer: a prospective multicenter study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 29(25):3457–3465, September 2011.
- [3] Martine Extermann, Ivette Boler, Richard R. Reich, Gary H. Lyman, Richard H. Brown, Joseph DeFelice, Richard M. Levine, Eric T. Lubiner, Pablo Reyes, Frederic J. Schreiber, and Lodovico Balducci. Predicting the risk of chemotherapy toxicity in older patients: the Chemotherapy Risk Assessment Scale for High-Age Patients (CRASH) score. *Cancer*, 118(13):3377–3386, July 2012.
- [4] National Cancer Institute. Cancer stat facts: Myeloma. Surveillance, Epidemiology, and End Results Program, 2025. Accessed 2026-03-21.
- [5] Penelope M. Webb and Susan J. Jordan. Epidemiology of epithelial ovarian cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 41:3–14, 2017.
- [6] Evangelos Terpos, Ioannis Ntanasis-Stathopoulos, Maria Gavriatopoulou, and Meletios A. Dimopoulos. Pathogenesis of bone disease in multiple myeloma: from bench to bedside. *Blood Cancer Journal*, 8(1):7, January 2018.
- [7] Meletios A. Dimopoulos, Giampaolo Merlini, Frank Bridoux, Nelson Leung, Joseph Mikhael, Simon J. Harrison, Efsthios Kastritis, Laurent Garderet, Alessandro Gozzetti, Niels W. C. J. van de Donk, Katja C. Weisel, Ashraf Z. Badros, Meral Beksac, Jens Hillengass, Mohamad Mohty, P. Joy Ho, Ioannis Ntanasis-Stathopoulos, Maria-Victoria Mateos, Paul Richardson, Joan Blade, Philippe Moreau, Jesus San-Miguel, Nikhil Munshi, S. Vincent Rajkumar, Brian G. M. Durie, Heinz Ludwig, and Evangelos Terpos. Management of multiple myeloma-related renal impairment: recommendations from the International Myeloma Working Group. *The Lancet Oncology*, 24(7):e293–e311, 2023.
- [8] Lois B. Travis, Eric J. Holowaty, Kerstin Bergfeldt, Charles F. Lynch, Betsy A. Kohler, Tomas Wiklund, Rochelle E. Curtis, Per Hall, Michael Andersson, Eero Pukkala, Jessica Sturgeon, and Marilyn Stovall. Risk of leukemia after platinum-based chemotherapy for ovarian cancer. *New England Journal of Medicine*, 340(5):351–357, 1999.
- [9] V. Tuninetti, J. A. Marín-Jiménez, G. Valabrega, and E. Ghisoni. Long-term outcomes of PARP inhibitors in ovarian cancer: survival, adverse events, and post-progression insights. *ESMO open*, 9(11):103984, November 2024.
- [10] Florent Malard, Paola Neri, Nizar J. Bahlis, Evangelos Terpos, Nour Moukalled, Vania T. M. Hungria, Salomon Manier, and Mohamad Mohty. Multiple myeloma. *Nature Reviews Disease Primers*, 10(1), June 2024.
- [11] Arti Hurria, Supriya Mohile, Ajeet Gajra, Heidi Klepin, Hyman Muss, Andrew Chapman, Tao Feng, David Smith, Can-Lan Sun, Nienke De Glas, Harvey Jay Cohen, Vani Katheria, Caroline Doan, Laura Zavala, Abraham Levi, Chie Akiba, and William P. Tew. Validation of a Prediction Tool for Chemotherapy Toxicity in Older Adults With Cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 34(20):2366–2371, July 2016.
- [12] Maxime Frelaut, Elena Paillaud, Guillaume Beinse, Anne-Laure Scain, Stéphane Culine, Christophe Tournigand, Johanne Poisson, Sylvie Bastuji-Garin, Florence Canoui-Poitine, and Philippe Caillet. External validity of two scores for predicting the risk of chemotherapy toxicity among older patients with solid tumors: Results from the ELCAPA prospective cohort. *The Oncologist*, 28(6):e341–e349, 2023.
- [13] Erin B. Moth, Belinda E. Kiely, Natalie Stefanic, Vasikaran Naganathan, Andrew Martin, Peter Grimison, Martin R. Stockler, Philip Beale, and Prunella Blinman. Predicting chemotherapy toxicity in older adults: Comparing the predictive value of the CARG toxicity score with oncologists’ estimates of toxicity based on clinical judgement. *Journal of Geriatric Oncology*, 10(2):202–209, 2019.

- [14] Zeshan Hussain, Edward De Brouwer, Rebecca Boiarsky, Sama Setty, Neeraj Gupta, Guohui Liu, Cong Li, Jaydeep Srimani, Jacob Zhang, Rich Labotka, and David Sontag. Joint AI-driven event prediction and longitudinal modeling in newly diagnosed and relapsed multiple myeloma. *npj Digital Medicine*, 7(1):200, July 2024.
- [15] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T Teo, and Richard J B Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024.
- [16] Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, September 2025.
- [17] Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025.
- [18] Chunyang Fu et al. A time series foundation model for cancer management. In *NeurIPS Time Series for Health Workshop*, 2025. OpenReview: BrkSpLTHe0.
- [19] Abu Sarwar Zamani, Abdelwahed Motwakel Eltayeb, Adel Al-luhayb, Md Mobin Akhtar, Rashid Ayub, Mohammed Abdelmonem Ahmed Abdelrahim, Sara Saadeldeen Ibrahim Mohamed, and Naved Ahmad. Application of machine learning in predicting cancer complications using longitudinal data: A systematic review and meta-analysis. *International Journal of Medical Informatics*, 208:106217, 2025.
- [20] Luoting Zhuang, Stephen H. Park, Steven J. Skates, Ashley E. Prosper, Denise R. Aberle, and William Hsu. Advancing precision oncology through modeling of longitudinal and multimodal data. *IEEE Reviews in Biomedical Engineering*, 19:182–200, 2026.
- [21] Laura Swinckels, Frank C. Bennis, Kirsten A. Ziesemer, Janneke F. M. Scheerman, Harmen Bijwaard, Ander de Keijzer, and Josef Jan Bruers. The Use of Deep Learning and Machine Learning on Longitudinal Electronic Health Records for the Early Detection and Prevention of Diseases: Scoping Review. *Journal of Medical Internet Research*, 26:e48320, August 2024.
- [22] Mark A. Perazella. Onco-nephrology: renal toxicities of chemotherapeutic agents. *Clinical journal of the American Society of Nephrology: CJASN*, 7(10):1713–1721, October 2012.
- [23] Noopur S. Raje, Elias Anaissie, Shaji K. Kumar, Sagar Lonial, Thomas Martin, Morie A. Gertz, et al. Consensus guidelines and recommendations for infection prevention in multiple myeloma: a report from the International Myeloma Working Group. *The Lancet Haematology*, 9(2):e143–e161, 2022.
- [24] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. Version 3.1.
- [25] Erik H. Rustad et al. The multiple myeloma research foundation compass study: Clinical and genomic characterization of multiple myeloma. *Nature Medicine*, 25:197–206, 2019.
- [26] Michael Chahin, Zachery Branham, Ashley Fox, Christian Leurinda, and Amany R. Keruakous. Clinical Considerations for Immunoparesis in Multiple Myeloma. *Cancers*, 14(9):2278, January 2022.
- [27] Antonio Palumbo, Jon Bladé, Mario Boccadoro, Carmela Paladino, Faith Davies, Meletios Dimopoulos, Anna Dmoszynska, Hermann Einsele, Philippe Moreau, Orhan Sezer, Andrew Spencer, Pieter Sonneveld, and Jesus San Miguel. How to Manage Neutropenia in Multiple Myeloma. *Clinical Lymphoma Myeloma and Leukemia*, 12(1):5–11, February 2012.
- [28] Elias K. Mai, Eva-Maria Haas, Stephan Lücke, Martin Löffrich, Christina Kunz, Maria Pritsch, Petra Knaup-Gregori, Marc S. Raab, Jana Schlenzka, Uta Bertsch, Jens Hillengass, and Hartmut Goldschmidt. A systematic classification of death causes in multiple myeloma. *Blood Cancer Journal*, 8(3):30, 2018.
- [29] Cecilie Hveding Blimark, Kristina Carlson, Christopher Day, Sigrun Einarsdottir, Gunnar Juliusson, Moshtak Karma, Dorota Knut-Bojanowska, Gunnar Larfors, Ingemar Turesson, Mariana Villegas-Scivetti, and Ingigerdur Sverrisdóttir. Risk of infections in multiple myeloma. A population-based study on 8672 multiple myeloma patients diagnosed 2008–2021 from the Swedish Myeloma Registry. *Haematologica*, July 2024.
- [30] Chandler S Callaway, Lila M Mouchantat, Benjamin G Bitler, and Andrea Bonetto. Mechanisms of Ovarian Cancer-Associated Cachexia. *Endocrinology*, 165(1):bqad176, November 2023.
- [31] Eduardo Gonzalez-Ochoa, Husam A Alqaisi, Gita Bhat, Nazlin Jivraj, and Stephanie Lheureux. Inoperable Bowel Obstruction in Ovarian Cancer: Prevalence, Impact and Management Challenges. *International Journal of Women’s Health*, 14:1849–1862, December 2022.
- [32] Gordon John Sampson Rustin, Ignace Vergote, Elizabeth Eisenhower, Eric Pujade-Lauraine, Michael Quinn, Tate Thigpen, Andreas du Bois, Gunnar Kristensen, Anders Jakobsen, Satoru Sagae, Kathryn Greven, Mahesh Parmar, Michael Friedlander, Andres Cervantes, and Jan Vermorken. Definitions for Response and Progression in Ovarian Cancer Clinical Trials Incorporating RECIST 1.1 and CA 125 Agreed by the Gynecological Cancer Intergroup (GCIg). *International Journal of Gynecological Cancer*, 21(2):419–423, January 2011.
- [33] Emine Bayraktar, Sisy Chen, Sara Corvigno, Jinsong Liu, and Anil K. Sood. Ovarian cancer metastasis: Looking beyond the surface. *Cancer Cell*, 42(10):1631–1636, October 2024.
- [34] Saad Z. Usmani, Christoph Heuck, Alan Mitchell, Jackie Szymonifka, Bijay Nair, Antje Hoering, Yazan Alsayed, Sarah Waheed, Sajjad Haider, Alejandro Restrepo, Frits Van Rhee, John Crowley, and Bart Barlogie. Extramedullary disease portends poor prognosis in multiple myeloma and is over-represented in high-risk disease even in the era of novel agents. *Haematologica*, 97(11):1761–1767, November 2012.
- [35] J. Rauw, R. A. Wells, A. Chesney, M. Reis, Liying Zhang, and Rena Buckstein. Validation of a scoring system to establish the probability of myelodysplastic syndrome in patients with

- unexplained cytopenias or macrocytosis. *Leukemia Research*, 35(10):1335–1338, October 2011.
- [36] Anna Galli, Gabriele Todisco, Eulalia Catamo, Cinzia Sala, Chiara Elena, Sara Pozzi, Elisa Bono, Virginia Valeria Ferretti, Ettore Rizzo, Elisabetta Molteni, Silvia Zibellini, Martina Sarchi, Emanuela Boveri, Jacqueline Ferrari, Nicolas Fiorelli, Clara Camaschella, Paolo Gasparini, Daniela Toniolo, Mario Cazzola, and Luca Malcovati. Relationship between clone metrics and clinical outcome in clonal cytopenia. *Blood*, 138(11):965–976, September 2021.
- [37] Xiurong Cai, Robert L. Bowman, and Jennifer J. Trowbridge. Clonal hematopoiesis in myeloid malignancies and solid tumors. *Nature Cancer*, 6(7):1133–1144, July 2025.
- [38] Adam G. Tabák, Christian Herder, Wolfgang Rathmann, Eric J. Brunner, and Mika Kivimäki. Prediabetes: a high-risk state for diabetes development. *The Lancet*, 379(9833):2279–2290, 2012.
- [39] Donald M. Lloyd-Jones and John T. Wilkins. Cardiovascular Risk Assessment and Prevention Across the Life Course. *JACC*, 81(7):633–635, February 2023.
- [40] K. V. Patel, L. Ferrucci, W. B. Ershler, D. L. Longo, and J. M. Guralnik. Red blood cell distribution width and the risk of death in middle-aged and older adults. *Arch Intern Med*, 169:515–523, 2009.
- [41] Marcio Nucci and Elias Anaissie. Infections in patients with multiple myeloma in the era of high-dose therapy and novel agents. *Clinical Infectious Diseases*, 49(8):1211–1225, 2009.
- [42] Takaya Okawa, Tomohiro Mizuno, Shogo Hanabusa, Takeshi Ikeda, Fumihiro Mizokami, Takenao Koseki, Kazuo Takahashi, Yukio Yuzawa, Naotake Tsuboi, Shigeki Yamada, and Yoshitaka Kameya. Prediction model of acute kidney injury induced by cisplatin in older adults using a machine learning algorithm. *PLoS One*, 17(1):e0262021, 2022.
- [43] Bum-Joo Cho, Kyoung Min Kim, Sanchir-Erdene Bilegsaikhan, and Yong Joon Suh. Machine learning improves the prediction of febrile neutropenia in Korean inpatients undergoing chemotherapy for breast cancer. *Scientific Reports*, 10(1):14803, 2020.
- [44] Ryuichiro Yagi, Shinichi Goto, Yoshinori Katsumata, Calum A. MacRae, and Rahul C. Deo. Artificial intelligence-enabled prediction of chemotherapy-induced cardiotoxicity from baseline electrocardiograms. *Nature Communications*, 15:2536, 2024.
- [45] Lachelle D. Weeks, Abhishek Niroula, Donna Neuberg, Waihay Wong, R. Coleman Lindsley, Marlise Luskin, Nancy Berliner, Richard M. Stone, Daniel J. DeAngelo, Robert Soiffer, Md Mesbah Uddin, Gabriel Griffin, Caitlyn Vlasschaert, Christopher J. Gibson, Siddhartha Jaiswal, Alexander G. Bick, Luca Malcovati, Pradeep Natarajan, and Benjamin L. Ebert. Prediction of risk for myeloid malignancy in clonal hematopoiesis. *NEJM Evidence*, 2(5), 2023.
- [46] Muxin Gu, Sruthi Cheloor Kovilakam, William G. Dunn, Ludovica Marando, Clea Barcena, Irina Mohorianu, Alexandra Smith, Siddhartha P. Kar, Margarete A. Fabre, Moritz Gerstung, Catherine A. Cargo, Luca Malcovati, Pedro M. Quiros, and George S. Vassiliou. Multiparameter prediction of myeloid neoplasia risk. *Nature Genetics*, 55(9):1523–1530, 2023.
- [47] Irenaeus C. C. Chan, Pu Zhang, Xiangyu Pan, Cynthia Castro, Nina Fox, Alexander M. Lewis, Kenyon Weis, Adriana Cuibus, Steven Tittley, Giulia Petrone, J. Scott Beeler, Duc Tran, Griffen Mustion, Catrina Fronick, Konrad H. Stopsack, Carlos Cruchaga, Omar Abdel-Wahab, and Kelly L. Bolton. CDK4/6 inhibition mitigates chemotherapy-induced expansion of TP53-mutant clonal hematopoiesis. *Nature Genetics*, 58(3):582–592, March 2026.
- [48] Nathalie Jetté, Hude Quan, Brenda Hemmelgarn, Saskia Drosler, Christina Maass, Lori Moskal, Wansa Paoiin, Vijaya Sundararajan, Song Gao, Robert Jakob, Bedihran Ustün, William A. Ghali, and IMECCHI Investigators. The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Medical Care*, 48(12):1105–1110, December 2010.
- [49] Reinhard Busse, Alexander Geissler, Wilm Quentin, and Miriam Wiley. *Diagnosis-Related Groups in Europe - Moving Towards Transparency, Efficiency and Quality in Hospitals*. 11 2011.
- [50] Pierre-Marie Morice, Alexandra Leary, Charles Dolladille, Basile Chrétien, Laurent Poulain, Antonio González-Martín, Kathleen Moore, Eileen Mary O’Reilly, Isabelle Ray-Coquard, and Joachim Alexandre. Myelodysplastic syndrome and acute myeloid leukaemia in patients treated with PARP inhibitors: a safety meta-analysis of randomised controlled trials and a retrospective study of the WHO pharmacovigilance database. *The Lancet Haematology*, 8(2):e115–e125, 2021.
- [51] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [52] S. Vincent Rajkumar. Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *American Journal of Hematology*, 97(8):1086–1107, May 2022.
- [53] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.
- [54] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.
- [55] E. Stephan et al. The connect mm registry: A large u.s. observational cohort study of patients with newly diagnosed multiple myeloma. *Clinical Lymphoma Myeloma and Leukemia*, 19(9):548–557, 2019.
- [56] N. Howlader et al. Overview of the seer program. *Journal of the National Cancer Institute Monographs*, 2014(49):1–10, 2014.

- [57] S. Vincent Rajkumar, Jean-Luc Harousseau, Brian Durie, Kenneth C. Anderson, Meletios Dimopoulos, Robert Kyle, Joan Blade, Paul Richardson, Robert Orlowski, David Siegel, Sundar Jagannath, Thierry Facon, Hervé Avet-Loiseau, Sagar Lonial, Antonio Palumbo, Jeffrey Zonder, Heinz Ludwig, David Vesole, Orhan Sezer, Nikhil C. Munshi, Jesus San Miguel, and on behalf of the International Myeloma Workshop Consensus Panel 1. Consensus recommendations for the uniform reporting of clinical trials: report of the international myeloma workshop consensus panel 1. *Blood*, 117(18):4691–4695, 05 2011.
- [58] Casper Hansen. casperhansen/llama-3.3-70b-instruct-awq. <https://huggingface.co/casperhansen/llama-3.3-70b-instruct-awq>, 2026. Accessed: 2026-01-17.

Extended Data

Table 1. Comparison of multiple myeloma datasets by available clinical content and data granularity.

Dataset	Comorbidities	Treatment	Medications	Labs / BM / Imaging / Pathology	Patients
MMRF CoMMpass [25]	Limited	Detailed	MM-specific only	Labs, BM, cytogenetics	1,143
Connect MM Registry [55]	Moderate	High-level	MM-specific only	Baseline labs	3,011
SEER [56]	None	Initial only, coded	None	None	population registry
This study	Detailed, narrative	Longitudinal, narrative	Complete	Comprehensive, narrative	2,328

Table 2. Laboratory parameters included in the analysis and their MIMIC-IV and CoMMpass MMRF equivalents. Each of the 39 routinely tested analytes used as model inputs was matched to its corresponding entry in the MIMIC-IV d_labitems table by clinical equivalence. Two analytes (plateletcrit and platelet distribution width) have no corresponding item in MIMIC-IV and were treated as missing inputs to the imputation model throughout the external cohort. The CoMMpass column lists the corresponding analyte code in the MMRF CoMMpass dataset where available.

Category	Laboratory Parameter	MIMIC itemid	CoMMpass
Complete Blood Count (CBC)	Basophils	51146	—
	Eosinophils	51200	—
	Hematocrit (Hct)	51221	—
	Hemoglobin (Hb)	51222	HGB
	Lymphocytes	51244	—
	Mean Corpuscular Hemoglobin	51248	—
	Mean Corpuscular Hem. Conc.	51249	—
	Mean Corpuscular Volume	51250	—
	Monocytes	51254	—
	Mean Platelet Volume	52142	—
	Neutrophils	51256	NEUT
	Plateletcrit	—	—
	Platelet Distribution Width	—	—
	Platelet Count	51265	PLT
	Red Blood Cell Count (RBC)	51279	—
	Red Cell Distribution Width	51277	—
	White Blood Cell Count (WBC)	51301	WBC
Electrolytes & Chemistry	Blood Urea Nitrogen	51006	BUN
	Creatinine	50912	CREAT
	Calcium (Ca ²⁺)	50893	CA
	Blood Glucose	50931	GLU
	Potassium (K ⁺)	50971	—
	Sodium (Na ⁺)	50983	—
	Anion Gap	50868	—
	Uric Acid	51007	—
Kidney Function	Estim. GFR (CKD-EPI)	53161	—
	Estim. GFR (MDRD Formula)	50920	—
Liver Function	Serum Albumin	50862	ALB
	Alkaline Phosphatase (ALP)	50863	—
	Alanine Aminotransferase (ALT)	50861	—
	Aspartate Aminotransferase (AST)	50878	—
	Gamma-Glutamyl Transferase	50927	—
	Total Bilirubin	50885	—
	Total Protein	50976	TP
Inflammation	C-Reactive Protein	50889	CRP
Coagulation	International Normalized Ratio	51237	—
	Prothrombin Time (PT)	51274	—
	Activ. Partial Throm. Time	51275	—
Thyroid Function	Thyroid-Stimulating Hormone	50993	—

Table 3. Endpoint label crosswalk from ICD-10-GM training labels to external-cohort definitions. For each ICD-10-GM endpoint used in training, the table lists the ICD-10-CM target codes used for MIMIC-IV validation and the corresponding MMRF CoMMpass concept used for CoMMpass validation. GM-specific axes not preserved in ICD-10-CM (NYHA class, coronary vessel count, decompensation status, KDIGO stage) were dropped during mapping and instead a closest match of available codes was applied to enable appropriate group-wise comparisons.

ICD-10-GM	GM description	ICD-10-CM (MIMIC-IV)	MMRF CoMMpass concept
A04.7	C. difficile enterocolitis	A04.7, A04.71, A04.72	C. difficile infection
A04.70	C. difficile enterocolitis	A04.7, A04.71, A04.72	C. difficile infection
A09	Gastroenteritis, infectious	A09	Infectious gastroenteritis
A09.0	Gastroenteritis, infectious	A09	Infectious gastroenteritis
A09.9	Gastroenteritis, unspecified	A09	Infectious gastroenteritis
A41.1	Sepsis, Staphylococcus	A41.1	Sepsis
A41.9	Sepsis, unspecified	A41.9	Sepsis
A41.51	Sepsis, E. coli	A41.51	Sepsis
A41.58	Sepsis, other Gram-negative	A41.59	Sepsis
B00.9	Herpesviral infection, unspecified	B00.9	Herpes simplex
B02.2	Zoster w/ nervous-system involvement	B02.2, B02.21, B02.29	Herpes zoster
B02.9	Herpes zoster, uncomplicated	B02.9	Herpes zoster
B18.1	Chronic viral hepatitis B	B18.1	Chronic hepatitis B/C
B18.2	Chronic viral hepatitis C	B18.2	Chronic hepatitis B/C
B24	HIV disease, unspecified	B20, B24	[unmappable]
B25.88	Other cytomegaloviral disease	B25.8	[unmappable]
B27.0	Infectious mononucleosis, EBV	B27.0, B27.00, B27.01, B27.02, B27.09	EBV mononucleosis
B35.1	Tinea unguium	B35.1	Onychomycosis / Fungal infection
B37.0	Candidal stomatitis	B37.0	Candidiasis / Fungal infection
B37.88	Candidiasis, other sites	B37.8, B37.81, B37.82, B37.83, B37.84, B37.89	Candidiasis / Fungal infection
B44.1	Pulmonary aspergillosis	B44.1	Aspergillosis / Fungal infection
B47.9	Mycetoma, unspecified	B47.9	Fungal infection
B95.2	Streptococcus grp D / Enterococcus	B95.2	Other bacterial infection
B95.6	Staphylococcus aureus	B95.6, B95.61, B95.62	Strep./staph. bacteraemia
B95.7	Other staphylococci	B95.7	Strep./staph. bacteraemia
B95.8	Unspecified staphylococci	B95.8	Strep./staph. bacteraemia
B95.48	Other specified streptococci	B95.4	Strep./staph. bacteraemia
B95.90	Other gram-positive aerobes	B96.8, B96.89	Strep./staph. bacteraemia
B96.2	E. coli / Enterobacterales	B96.2, B96.20, B96.21, B96.22, B96.23, B96.29	Other bacterial infection
B96.5	Pseudomonas	B96.5	Other bacterial infection
B96.6	Bacteroides fragilis	B96.6	Other bacterial infection
B96.8	Other specified bacterial agents	B96.8, B96.81, B96.82, B96.83, B96.89	Other bacterial infection
B99	Other infectious disease, unspec.	B99, B99.8, B99.9	Unspecified infection
C73	Malig. neoplasm, thyroid gland	C73	Thyroid cancer (SPM if post-MM)
C77.0	Sec. malig., head/neck LN	C77.0	[construct shift]
C77.1	Sec. malig., thoracic LN	C77.1	[construct shift]
C77.2	Sec. malig., intra-abdominal LN	C77.2	[construct shift]
C77.3	Sec. malig., axillary/UE LN	C77.3	[construct shift]
C77.4	Sec. malig., inguinal/LE LN	C77.4	[construct shift]
C77.5	Sec. malig., intrapelvic LN	C77.5	[construct shift]
C77.8	Sec. malig., LN multiple regions	C77.8	[construct shift]
C77.9	Sec. malig., LN unspecified	C77.9	[construct shift]
C78.0	Sec. malig., lung	C78.0, C78.00, C78.01, C78.02	[construct shift]
C78.1	Sec. malig. neoplasm, mediastinum	C78.1	[construct shift]
C78.2	Sec. malig., pleura	C78.2	[construct shift]
C78.4	Sec. malig., small intestine	C78.4	[construct shift]
C78.5	Sec. malig., large intestine	C78.5	[construct shift]
C78.6	Sec. malig., peritoneum	C78.6	[construct shift]
C78.7	Sec. malig., liver	C78.7	[construct shift]
C78.8	Sec. malig., other digestive	C78.8, C78.80, C78.89	[construct shift]

Continued on next page

ICD-10-GM	GM description	ICD-10-CM (MIMIC-IV)	MMRF CoMMpass concept
C79.2	Sec. malig., skin	C79.2	[construct shift]
C79.3	Sec. malig., brain/meninges	C79.3, C79.31, C79.32	[construct shift]
C79.4	Sec. malig., nervous system	C79.4, C79.40, C79.49	[construct shift]
C79.5	Sec. malig., bone/marrow	C79.5, C79.51, C79.52	[construct shift]
C79.6	Sec. malig., ovary	C79.6, C79.60, C79.61, C79.62, C79.63	[construct shift]
C79.7	Sec. malig., adrenal gland	C79.7, C79.70, C79.71, C79.72	[construct shift]
C79.9	Sec. malig., unspecified site	C79.9	[construct shift]
C79.82	Sec. malig., genital organs	C79.82	[construct shift]
C79.88	Sec. malig., other sites	C79.89	[construct shift]
D46.2	Refractory anaemia w/ excess blasts	D46.2, D46.20, D46.21, D46.22	[unmappable]
D46.7	Other myelodysplastic syndromes	D46.Z	[unmappable]
D46.9	Myelodysplastic syndrome, unspec.	D46.9	[unmappable]
E03.8	Other hypothyroidism	E03.8	Hypothyroidism
E03.9	Hypothyroidism, unspecified	E03.9	Hypothyroidism
E04.2	Nontoxic multinodular goitre	E04.2	Goitre/thyroid nodule
E04.9	Nontoxic goitre, unspecified	E04.9	Goitre/thyroid nodule
E05.0	Hyperthyroidism, Graves'	E05.00, E05.01	Hyperthyroidism
E05.9	Hyperthyroidism, unspecified	E05.90, E05.91	Hyperthyroidism
E06.3	Autoimmune thyroiditis (Hashimoto)	E06.3	Thyroid disorder
E11.20	T2DM with kidney complications	E11.21, E11.22, E11.29	Diabetes / Chronic kidney disease
E11.90	T2DM w/o complications	E11.9	Diabetes
E11.91	T2DM w/o complications (decompensated)	E11.9	Diabetes
E13.90	Other specified diabetes mellitus	E13.9	Diabetes
F32.0	Mild depressive episode	F32.0	Depression
F32.1	Moderate depressive episode	F32.1	Depression
F32.2	Severe depression w/o psychosis	F32.2	Depression
F32.8	Other depressive episodes	F32.89	Depression
F32.9	Depressive episode, unspecified	F32.9, F32.A	Depression
F33.1	Recurrent depression, moderate	F33.1	Depression
F33.2	Recurrent depression, severe	F33.2	Depression
I10	Essential hypertension (parent)	I10	Hypertension
I10.00	Essential hypertension, benign	I10	Hypertension
I10.01	Malignant essential hypertension	I10	Hypertension
I10.90	Essential hypertension, no crisis	I10	Hypertension
I10.91	Essential hypertension, w/ crisis	I10	Hypertension
I11.00	Hypertensive heart disease, w/ HF	I11.0	Hypertension / Heart failure
I11.90	Hypertensive heart disease, no HF	I11.9	Hypertension
I12.00	Hypertensive renal disease, w/ failure	I12.0, I12.9	Hypertension / Chronic kidney disease
I15.10	Renovascular hypertension	I15.0	Hypertension
I20.0	Unstable angina	I20.0	Angina
I20.8	Other angina pectoris	I20.8	Angina
I21.4	Acute subendocardial MI (NSTEMI)	I21.4	Myocardial infarction
I25.9	Chronic ischaemic HD, unspecified	I25.9	Coronary artery disease
I25.11	Atherosclerotic HD, 1-vessel	I25.10, I25.11x	Coronary artery disease / Angina
I25.12	Atherosclerotic HD, 2-vessel	I25.10, I25.11x	Coronary artery disease
I25.13	Atherosclerotic HD, 3-vessel	I25.10, I25.11x	Coronary artery disease
I25.19	Atherosclerotic HD, unspecified	I25.10, I25.11x	Coronary artery disease
I25.22	Old myocardial infarction	I25.2	Coronary artery disease
I42.0	Dilated cardiomyopathy	I42.0	Cardiomyopathy
I48.0	Atrial fibrillation, paroxysmal	I48.0	Atrial fibrillation/flutter
I48.1	Atrial fibrillation, persistent	I48.1, I48.11, I48.19	Atrial fibrillation/flutter
I48.2	Atrial fibrillation, chronic	I48.2, I48.20, I48.21, I48.91	Atrial fibrillation/flutter
I48.9	Atrial fibrillation, unspecified	I48.9, I48.91, I48.92	Atrial fibrillation/flutter
I48.10	Persistent AF (non-standard)	I48.1, I48.19	Atrial fibrillation/flutter
I48.11	Persistent AF, longstanding (non-std)	I48.1, I48.11	Atrial fibrillation/flutter
I48.19	Persistent AF, other (non-std)	I48.1, I48.19	Atrial fibrillation/flutter

Continued on next page

ICD-10-GM	GM description	ICD-10-CM (MIMIC-IV)	MMRF CoMMpass concept
I50.01	Secondary right heart failure	I50.81, I50.9	Heart failure
I50.9	Heart failure, unspecified	I50.9	Heart failure
I50.12	Left heart failure, NYHA II	I50.1, I50.2x, I50.3x, I50.4x	Heart failure
I50.13	Left heart failure, NYHA III	I50.1, I50.2x, I50.3x, I50.4x	Heart failure
I50.14	Left heart failure, NYHA IV	I50.1, I50.2x, I50.3x, I50.4x	Heart failure
I80.1	Phlebitis/thrombosis, femoral vein	I80.1, I82	Deep/superficial vein thrombosis
I80.2	Phlebitis, deep LE veins	I80.2, I82	Deep/superficial vein thrombosis
I80.3	Phlebitis, lower extremity (incl. DVT NOS)	I80.3, I82	Deep/superficial vein thrombosis
I80.8	Phlebitis, other localisation	I80.8, I80.89	Deep/superficial vein thrombosis
I80.28	Phlebitis, other deep LE veins	I80.29, I82	Deep/superficial vein thrombosis
I80.80	Phlebitis, superficial UE veins	I80.8, I80.89	Deep/superficial vein thrombosis
I80.81	Phlebitis, deep UE veins	I80.8, I80.89, I82.A1	Deep/superficial vein thrombosis
I80.88	Phlebitis, other localisation	I80.8, I80.89, I80.9, I82.90, I82.91	Deep/superficial vein thrombosis
I82.9	Embolism/thrombosis, unspec. veins	I82.90, I82.91	Embolism/thrombosis
J96.0	Acute respiratory failure	J96.0, J96.00, J96.01, J96.02	Respiratory failure
J96.00	Acute respiratory failure, type I	J96.01	Respiratory failure
J96.9	Respiratory failure, unspecified	J96.9, J96.90, J96.91, J96.92	Respiratory failure
J96.09	Acute respiratory failure, unspec.	J96.00	Respiratory failure
K92.2	GI haemorrhage, unspecified	K92.2	GI haemorrhage
M80.88	Osteoporosis w/ pathologic fracture	M80.88	Pathologic/compression fracture / Osteoporosis/osteopenia
M81.89	Other osteoporosis w/o fracture	M81.8	Osteoporosis/osteopenia
M81.99	Osteoporosis, unspecified	M81.9	Osteoporosis/osteopenia
M89.50	Osteolysis, multi-site / unspec.	M89.50, M89.59	Lytic bone lesion
M89.55	Osteolysis, femur	M89.55, M89.551, M89.552, M89.559	Lytic bone lesion
M89.58	Osteolysis, other site	M89.58	Lytic bone lesion
M89.59	Osteolysis, unspecified site	M89.50, M89.59	Lytic bone lesion
N17.8	Other acute kidney failure	N17.8	Acute kidney injury
N17.9	AKI, unspecified	N17.9	Acute kidney injury
N17.91	AKI, stage 1	N17.9	Acute kidney injury
N17.92	AKI, KDIGO stage 2	N17.9	Acute kidney injury
N17.93	AKI, stage 3	N17.9	Acute kidney injury
N17.99	AKI, stage unspecified	N17.9	Acute kidney injury
N18.0	End-stage renal disease (legacy)	N18.6, N18.5	Chronic kidney disease / Dialysis dependence
N18.2	CKD stage 2	N18.2	Chronic kidney disease
N18.3	CKD stage 3	N18.3, N18.30, N18.31, N18.32	Chronic kidney disease
N18.4	CKD stage 4	N18.4	Chronic kidney disease
N18.5	CKD stage 5 (incl. dialysis)	N18.5, N18.6	Chronic kidney disease
N18.9	CKD, unspecified	N18.9	Chronic kidney disease
N18.82	Other CKD (non-standard)	N18.9	Chronic kidney disease
N18.89	Other CKD, unspecified stage	N18.9	Chronic kidney disease
R73.9	Hyperglycaemia, unspecified	R73.9	Hyperglycaemia
R99	Ill-defined cause of mortality	R99	Death (vital status)
S32.01	Fracture, L1 vertebra	S32.01	Vertebral fracture
S42.3	Fracture, humerus shaft	S42.3	Upper-extremity fracture
S42.20	Fracture, proximal humerus, unspecified	S42.20	Upper-extremity fracture
S42.21	Fracture, surgical neck of humerus	S42.21	Upper-extremity fracture
S52.50	Distal radius fracture, unspecified	S52.50	Upper-extremity fracture
S52.51	Distal radius fracture (Colles)	S52.5	Upper-extremity fracture
S52.59	Other distal radius fractures	S52.5	Upper-extremity fracture
S72.00	Femoral neck fracture, unspecified	S72.0	Hip/femur fracture
S72.10	Pertrochanteric fracture	S72.1	Hip/femur fracture
S82.6	Fracture, lateral malleolus	S82.6	Lower-extremity fracture
S82.82	Fracture, other lower leg	S82.89	Lower-extremity fracture

Continued on next page

ICD-10-GM	GM description	ICD-10-CM (MIMIC-IV)	MMRF CoMMpass concept
Z99.1	Long-term respirator dependence	Z99.1, Z99.11	Respirator dependence (resp. failure proxy)
Z99.2	Dialysis dependence	Z99.2	Dialysis dependence
All-cause mortality	All-cause mortality	(hospital_expire_flag / dod)	Death (vital status)

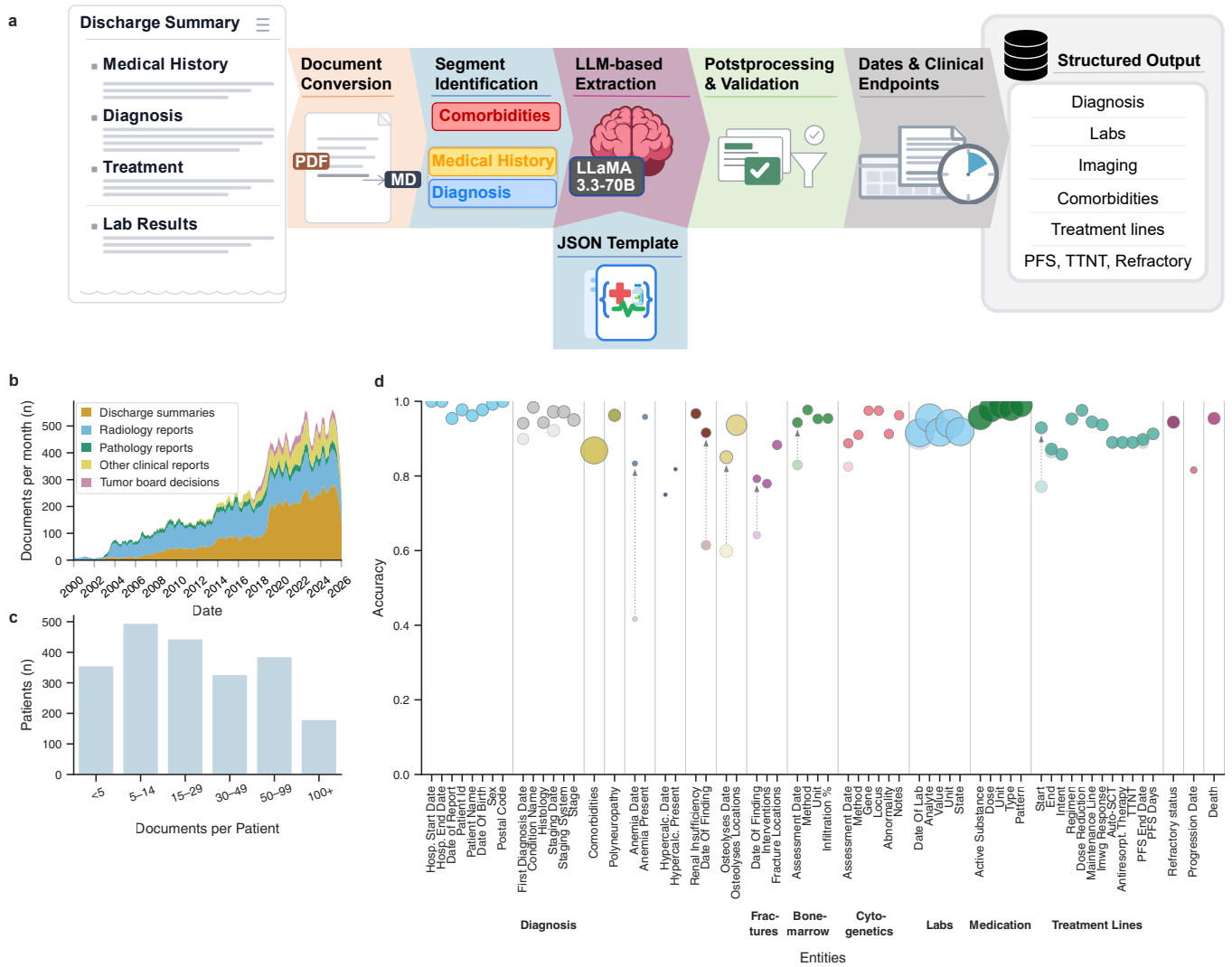


Figure 8. Automated extraction of structured clinical data from unstructured medical documents. **a**, Overview of the information extraction pipeline. Clinical documents are converted from PDF to markdown, segmented into sections (comorbidities, medical history, diagnosis, treatment, laboratory results), and processed through LLaMA 3.3-70B using domain-specific, JSON templates. Extracted entities undergo postprocessing, validation, and temporal alignment to produce structured representations spanning diagnosis, laboratory measurements, imaging, comorbidities, treatment lines, and clinical endpoints. **b**, Volume of clinical documents processed over time, stratified by document type ($n = 2,328$ patients, 2000–2026). Discharge summaries constitute the majority of records, with increasing document volumes over the observation period. **c**, Distribution of document counts per patient, illustrating the heterogeneity in clinical documentation density across the cohort. **d**, Field-level extraction accuracy (exact match) across 60 clinically relevant variables grouped by domain. Each point represents one extracted variable. Point size is proportional to the number of annotated instances (N). For date variables, arrows extending upward from the opaque data points to semi-transparent points indicate the change from strict to relaxed date matching, which permits a broader temporal window. Accuracy exceeds 0.90 for the majority of variables, with lower performance observed for date fields with few occurrences (e.g., anemia, hypercalcemia, fractures). Detailed per-variable results including relaxed accuracy metrics are provided in Table 4.

Supplementary Tables

Table 4. Field-level extraction accuracy across clinically relevant domains. Accuracy denotes exact-match agreement between automated extraction and expert annotation. *Accuracy (relaxed)* reports accuracy when allowing a broader temporal window for date matching. *N* indicates the total number of annotated instances evaluated per variable.

Clinical domain	Extracted variable	Accuracy	Accuracy (relaxed)	N
Patient demographics	Patient name	0.962	–	131
	Date of birth	0.977	0.977	131
	Sex	0.992	–	131
	Postal code	1.000	–	131
Hospitalization	Document generation date	0.954	0.954	131
	Hospital admission date	1.000	1.000	131
	Hospital discharge date	1.000	1.000	129
	Patient identifier	0.977	–	131
Diagnosis (structured)	Date of first diagnosis	0.899	0.941	119
	Primary diagnosis name	0.984	–	123
	Histological subtype	0.943	–	123
	Tumor staging date	0.922	0.972	141
	Tumor stage	0.950	–	141
	Staging system	0.972	–	141
Medication	Active substance	0.957	–	532
	Administration schedule pattern	0.989	–	458
	Administration schedule type	0.981	–	530
	Dose	0.977	–	520
	Unit	0.987	–	520
Laboratory measurements	Analyte	0.956	–	699
	Date of laboratory assessment	0.909	0.916	700
	Result status	0.919	–	701
	Unit	0.940	–	705
	Value	0.917	–	703
Bone marrow assessment	Assessment date	0.830	0.943	88
	Assessment method	0.977	–	87
	Measurement unit	0.953	–	85
	Infiltration %	0.954	–	87
Cytogenetic assessment	Assessment date	0.825	0.888	80
	Assessment method	0.910	–	78
	Affected gene	0.975	–	80
	Genomic locus	0.975	–	80
	Abnormality	0.913	–	80
	Notes	0.963	–	80
Treatment lines	Treatment start date	0.772	0.929	127
	Treatment end date	0.864	0.872	125
	Treatment intent	0.858	–	127
	Regimen	0.953	–	127
	Dose reduction	0.976	–	125
	Maintenance therapy	0.945	–	127
	IMWG response	0.937	–	127
	HDT with autologous SCT	0.890	–	127
	Antiresorptive therapy	0.890	–	127
	Time to next treatment	0.890	–	127
Outcomes	Progression-free survival end date	0.890	0.898	127
	Progression-free survival interval	0.913	–	126
	Refractory status	0.944	–	133
	Progression date	0.816	0.816	38
	Death date	0.954	–	131
Comorbidities	Diagnosis date and name	0.868	–	646
	Polyneuropathy present	0.963	–	134
Anemia	Anemia date	0.417	0.833	24
	Anemia present	0.958	–	24
Hypercalcemia	Hypercalcemia date	0.750	0.750	12
	Hypercalcemia present	0.818	–	11
Renal complications	Renal insufficiency date	0.614	0.916	83
	Renal insufficiency present	1.000	–	83

Continued on next page

Clinical domain	Extracted variable	Accuracy	Accuracy (relaxed)	N
Fractures	Fracture finding date	0.614	0.916	83
	Fracture locations	0.883	–	77
	Interventions	0.779	–	68
	Intervention date	0.642	0.792	53
Osteolyses	Date of diagnosis	0.599	0.850	147
	Lesion locations	0.936	–	377

Table 5. Field-level extraction accuracy across clinically relevant domains on the external dataset (document-level evaluation). Accuracy denotes exact-match agreement between automated extraction and expert annotation. *N* indicates the total number of annotated instances evaluated per variable.

Clinical domain	Extracted variable	Accuracy	N
Patient demographics	Patient name	1.000	13
	Date of birth	1.000	13
	Sex	1.000	13
	Postal code	0.846	13
Hospitalization	Document generation date	0.769	13
	Hospital admission date	0.846	13
	Hospital discharge date	0.846	13
	Patient identifier	1.000	9
Diagnosis (structured)	Date of first diagnosis	1.000	11
	Primary diagnosis name	1.000	10
	Histological subtype	1.000	10
	Tumor staging date	1.000	9
	Tumor stage	1.000	9
	Staging system	1.000	9
Medication	Active substance	0.895	171
	Administration schedule pattern	0.893	112
	Administration schedule type	0.856	174
	Dose	0.870	169
	Unit	0.870	169
Laboratory measurements	Analyte	0.894	283
	Date of laboratory assessment	0.879	282
	Result status	0.781	283
	Unit	0.883	282
	Value	0.855	282
Bone marrow assessment	Assessment date	1.000	5
	Assessment method	1.000	5
	Measurement unit	1.000	5
	Infiltration %	1.000	5
Treatment lines	Treatment start date	0.880	25
	Treatment end date	0.840	25
	Treatment intent	0.880	25
	Regimen	0.920	25
	Dose reduction	0.840	25
	Maintenance therapy	0.760	25
	IMWG response	0.960	25
	HDT with autologous SCT	0.960	25
	Antiresorptive therapy	0.680	25
	Time to next treatment	0.760	25
Outcomes	Progression-free survival end date	0.800	25
	Progression-free survival interval	0.800	25
	Progression date	0.900	20
	Death date	1.000	13
Comorbidities	Diagnosis date and name	0.961	77
	Polyneuropathy present	1.000	3
Anemia	Anemia date	0.333	9
	Anemia present	0.750	8
Hypercalcemia	Hypercalcemia present	0.923	13
Renal complications	Renal insufficiency date	0.333	3
	Renal insufficiency present	1.000	3
Fractures	Fracture finding date	0.938	16
	Fracture locations	1.000	23
	Interventions	1.000	16
Osteolyses	Date of diagnosis	0.800	5
	Lesion locations	0.957	23

Table 6. Predictive performance and prediction horizon across diagnostic endpoints ordered by prevalence. Model performance is reported using median area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals (CI) from repeated resampling for the Transformer. For XGBoost and logistic regression, only the median AUROC is reported. Median average precision (AP) is additionally reported for the Transformer. The prediction horizon denotes the median time in days from the end of the laboratory input window to the diagnosis event, with interquartile range (IQR). Evaluated on laboratory measurements sampled from randomly positioned 10–30 day intervals prior to diagnosis. Held-out test set comprising 1,579 sequences. Rows are ordered by decreasing number of positive cases (*N*). Endpoints with no positive cases in the test set are marked “—”.

Endpoint	Transformer				XGBoost	Log. Reg.	<i>N</i>	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
All-cause mortality	0.697	0.653	0.737	0.213	0.656	0.660	187	346	195–531.5
All-cause mortality									
B96.2	0.661	0.616	0.703	0.256	0.628	0.633	175	210	69–450.5
Escherichia coli									
I10.00	0.722	0.680	0.763	0.291	0.680	0.700	159	112	24–283
Essential hypertension, benign									
C78.6	0.888	0.862	0.913	0.471	0.882	0.870	117	97	20–255
Sec. malig. neoplasm, peritoneum									
B95.2	0.660	0.609	0.715	0.124	0.637	0.689	110	230.5	84–426
Streptococcus grp D / Enterococcus									
I10.90	0.768	0.723	0.807	0.176	0.659	0.735	97	139	43–330
Essential hypertension, unspec.									
E03.8	0.694	0.640	0.749	0.105	0.674	0.686	86	210	42.2–365
Other hypothyroidism									
M89.50	0.847	0.813	0.876	0.238	0.857	0.870	85	192	96–350
Osteolysis, multiple sites									
C78.2	0.858	0.825	0.886	0.208	0.774	0.797	74	216.5	86.8–390.2
Sec. malig. neoplasm, pleura									
N17.91	0.790	0.742	0.832	0.113	0.664	0.767	68	255	98.8–419.5
Acute kidney failure, stage 1									
I80.28	0.597	0.532	0.659	0.058	0.618	0.657	66	193.5	76.8–370
Phlebothrombosis, other									
A09.9	0.656	0.589	0.726	0.133	0.545	0.492	65	240	90–436
Gastroenteritis, unspecified									
C79.5	0.567	0.485	0.649	0.104	0.569	0.566	64	229	100–468.8
Sec. malig. neoplasm, bone/marrow									
C77.2	0.865	0.831	0.901	0.224	0.782	0.825	64	124.5	35–281.2
Sec. malig. neoplasm, intra-abd. LN									
I48.0	0.754	0.696	0.813	0.147	0.674	0.775	64	159	56–397.8
Paroxysmal atrial fibrillation									
B99	0.588	0.499	0.669	0.063	0.542	0.600	61	160	76–438
Other infectious disease, unspec.									
A41.9	0.742	0.666	0.810	0.097	0.590	0.737	58	202	60.8–429
Sepsis, unspecified									
E03.9	0.655	0.586	0.723	0.072	0.629	0.601	57	121	23–354
Hypothyroidism, unspecified									
E11.90	0.628	0.550	0.702	0.068	0.542	0.585	54	191.5	73.8–391
Type 2 diabetes, w/o complications									
N18.3	0.877	0.836	0.911	0.157	0.856	0.860	54	286	150–483.2
Chronic kidney disease, stage 3									
A04.70	0.770	0.690	0.839	0.137	0.699	0.667	52	154	83–287.5
Enterocolitis, C. difficile									
C79.88	0.756	0.688	0.813	0.088	0.742	0.758	49	210	55–349
Sec. malig. neoplasm, other sites									
C78.7	0.726	0.672	0.777	0.058	0.711	0.765	48	271.5	126.2–475
Sec. malig. neoplasm, liver									
B95.7	0.580	0.478	0.673	0.047	0.609	0.602	46	186	85.5–332.8
Other staphylococci									
J96.00	0.751	0.675	0.817	0.074	0.691	0.750	45	136	52–301
Acute respiratory failure									

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
B96.5 Pseudomonas	0.656	0.574	0.737	0.050	0.654	0.624	45	181	69–397
N18.4 Chronic kidney disease, stage 4	0.853	0.781	0.905	0.210	0.751	0.790	43	230	115–425.5
I10.01 Malignant essential hypertension	0.800	0.737	0.852	0.086	0.532	0.581	39	157	59.5–442
K92.2 GI haemorrhage, unspecified	0.687	0.603	0.766	0.046	0.520	0.595	38	215.5	141–386.8
N17.9 Acute kidney failure, unspecified	0.835	0.786	0.880	0.076	0.864	0.916	37	288	109–400
E04.2 Nontoxic multinodular goitre	0.686	0.599	0.771	0.040	0.575	0.589	35	313	134.5–504.5
N18.9 Chronic kidney disease, unspecified	0.854	0.782	0.916	0.108	0.838	0.809	31	230	130–318.5
M81.99 Osteoporosis, unspecified site	0.608	0.500	0.715	0.032	0.602	0.663	30	410.5	151–543.5
A04.7 Enterocolitis, C. difficile	0.841	0.776	0.899	0.075	0.820	0.902	29	370	142–463
B37.0 Candidal stomatitis	0.878	0.825	0.920	0.118	0.802	0.834	27	125	54.5–255
N17.99 Acute kidney failure, unspec. stage	0.709	0.620	0.805	0.049	0.647	0.689	27	105	50–292
I50.14 Left heart failure, NYHA IV	0.805	0.722	0.881	0.113	0.528	0.415	27	313	157–547.5
N18.5 Chronic kidney disease, stage 5	0.901	0.836	0.955	0.416	0.849	0.861	26	181	66–331.2
I50.9 Heart failure, unspecified	0.457	0.342	0.579	0.013	0.504	0.482	22	388	191–538.8
I48.10 Atrial flutter, unspecified	0.888	0.809	0.947	0.117	0.892	0.902	22	130.5	46.5–310.5
C77.3 Sec. malig. neoplasm, axillary LN	0.701	0.611	0.786	0.024	0.586	0.678	22	231.5	167–384.5
I25.11 Atherosclerotic HD, native coronary	0.729	0.639	0.813	0.036	0.522	0.660	21	133	44–335
I25.19 Atherosclerotic HD, other	0.498	0.367	0.644	0.022	0.530	0.522	21	206	113–413
C77.5 Sec. malig. neoplasm, intrapelvic LN	0.775	0.664	0.873	0.055	0.774	0.781	20	79	22–214.8
B37.88 Other candidiasis	0.685	0.571	0.785	0.022	0.729	0.669	20	211	112.5–283.5
D46.9 Myelodysplastic syndrome, unspec.	0.715	0.585	0.833	0.067	0.581	0.581	20	368	326.2–459
I11.00 Hypertensive heart disease	0.844	0.759	0.921	0.154	0.515	0.743	20	173	23.8–513
B96.8 Other specified bacterial agents	0.533	0.414	0.643	0.013	0.380	0.361	20	225	135.5–356
A41.58 Other Gram-negative sepsis	0.589	0.449	0.721	0.020	0.584	0.559	20	407	100.8–582
F32.9 Depressive episode, unspecified	0.529	0.407	0.659	0.014	0.563	0.563	19	167	101.5–360
I50.13 Left heart failure, NYHA III	0.843	0.743	0.929	0.104	0.603	0.733	19	165	47–463.5
Z99.1 Dependence on respirator	0.770	0.682	0.846	0.027	0.739	0.691	19	316	134–436

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
B95.8	0.906	0.818	0.964	0.140	0.802	0.848	18	112	73.2–151.2
Other streptococci/staphylococci									
B95.48	0.793	0.726	0.862	0.030	0.603	0.604	18	387.5	233–514.5
Other streptococci									
J96.09	0.704	0.566	0.822	0.029	0.711	0.801	18	284.5	191.5–361.5
Acute respiratory failure, other									
M89.58	0.855	0.791	0.911	0.059	0.737	0.781	17	198	31–457
Osteolysis, other site									
I48.9	0.801	0.680	0.910	0.066	0.707	0.813	17	252	47–509
Atrial fibrillation, unspecified									
Z99.2	0.807	0.650	0.936	0.267	0.949	0.906	17	517	282–578
Dependence on renal dialysis									
N17.8	0.791	0.706	0.865	0.026	0.882	0.846	17	338	105–440
Other acute kidney failure									
E04.9	0.551	0.425	0.669	0.012	0.555	0.505	16	229	106.8–366
Nontoxic goitre, unspecified									
C78.8	0.752	0.676	0.822	0.020	0.566	0.475	16	115.5	45–290.5
Sec. malig. neoplasm, other digest.									
F32.1	0.586	0.413	0.744	0.016	0.601	0.560	14	232	68–390.5
Moderate depressive episode									
I50.12	0.806	0.693	0.907	0.056	0.768	0.764	14	40	10.2–94.8
Left heart failure, NYHA II									
C78.0	0.736	0.560	0.872	0.036	0.839	0.736	14	160	54.8–253.2
Sec. malig. neoplasm, lung									
B02.9	0.692	0.579	0.797	0.027	0.441	0.276	14	273.5	179.5–455.8
Herpes zoster, uncomplicated									
C77.9	0.691	0.586	0.804	0.084	0.542	0.644	14	125.5	34.5–286.5
Sec. malig. neoplasm, LN unspec.									
I50.01	0.911	0.868	0.945	0.052	0.725	0.854	14	185.5	48.8–378.5
Right heart failure									
M80.88	0.654	0.523	0.787	0.014	0.414	0.718	14	313	136–602.8
Osteoporosis w/ fracture, other site									
R99	0.769	0.664	0.860	0.022	0.677	0.811	14	289	204–372
Ill-defined / unknown cause of death									
I25.9	0.885	0.844	0.924	0.036	0.849	0.925	13	170	10–493
Chronic ischaemic HD, unspec.									
E11.91	0.875	0.760	0.947	0.064	0.762	0.621	13	99	35–239
Type 2 diabetes, w/o complications (decompensated)									
E06.3	0.550	0.371	0.728	0.021	0.638	0.655	13	237	75–364
Autoimmune thyroiditis									
B95.6	0.703	0.532	0.842	0.024	0.694	0.633	13	152	115–259
Staphylococcus aureus									
A41.1	0.914	0.840	0.971	0.163	0.816	0.891	13	236	185–477
Sepsis, other Staphylococcus									
I48.1	0.694	0.521	0.844	0.021	0.709	0.662	13	488	15–565
Persistent atrial fibrillation									
N17.93	0.762	0.627	0.895	0.080	0.711	0.736	12	175	97–313.5
Acute kidney failure, stage 3									
N18.2	0.830	0.712	0.918	0.030	0.740	0.792	12	500	328.8–555
Chronic kidney disease, stage 2									
R73.9	0.790	0.723	0.858	0.033	0.640	0.664	12	223.5	112.5–407.5
Hyperglycaemia, unspecified									
I12.00	0.788	0.670	0.886	0.030	0.685	0.771	12	445	279.5–547.5
Hypertensive renal dis. w/ failure									

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
E05.9 Thyrototoxicosis, unspecified	0.681	0.520	0.819	0.016	0.671	0.718	12	199	65–505
F32.2 Severe depressive episode, no psychosis	0.696	0.529	0.850	0.018	0.574	0.754	11	38	13–130.5
I25.22 Old myocardial infarction, other	0.804	0.672	0.919	0.042	0.662	0.689	11	74	6–317.5
B44.1 Pulmonary aspergillosis	0.854	0.750	0.943	0.052	0.858	0.841	11	196	59.5–318
I21.4 Acute subendocardial MI	0.714	0.615	0.807	0.011	0.527	0.579	11	243	66.5–448.5
C79.2 Sec. malig. neoplasm, skin	0.679	0.509	0.839	0.023	0.619	0.527	11	285	212–568.5
C77.1 Sec. malig. neoplasm, thoracic LN	0.784	0.705	0.860	0.019	0.617	0.753	11	195	47.5–407
B96.6 Bacteroides fragilis	0.650	0.407	0.846	0.019	0.559	0.590	10	220.5	119.2–372.2
C79.3 Sec. malig. neoplasm, brain	0.458	0.321	0.604	0.006	0.676	0.807	10	213.5	126.5–269.8
A09.0 Gastroenteritis, infectious	0.689	0.481	0.854	0.012	0.339	0.480	10	144	101.8–430.5
C77.8 Sec. malig. neoplasm, LN multiple	0.848	0.735	0.940	0.028	0.789	0.788	10	318.5	241.8–456.2
F33.1 Recurrent depressive disorder, mod.	0.702	0.561	0.819	0.011	0.336	0.304	10	130	31.8–300.2
M89.55 Osteolysis, femur	0.825	0.765	0.901	0.121	0.664	0.724	9	223	91–251
I10.91 Hypertensive crisis	0.841	0.746	0.922	0.020	0.551	0.418	9	195	87–307
I80.1 Phlebitis/thromb., femoral vein	0.633	0.464	0.792	0.008	0.701	0.701	9	76	41–154
S82.6 Fracture, lateral malleolus	0.333	0.193	0.474	0.005	0.154	0.480	9	535	438–577
S32.01 Fracture, lumbar vertebra L1	0.873	0.751	0.981	0.065	0.857	0.873	9	484	307–595
I20.0 Unstable angina	0.703	0.473	0.883	0.018	0.706	0.672	8	64	19.8–105.2
I82.9 Venous embolism, unspecified	0.386	0.099	0.662	0.007	0.520	0.475	8	205	111–317.5
I15.10 Secondary hypertension, renal	0.674	0.438	0.912	0.080	0.932	0.917	8	536.5	344.2–608
I80.80 Phlebitis, superficial UE veins	0.606	0.521	0.701	0.006	0.479	0.523	8	238	85.2–524
C78.5 Sec. malig. neoplasm, large intestine	0.909	0.818	0.964	0.034	0.718	0.728	8	15.5	7–25.5
S72.10 Pertrochanteric femoral fracture	0.406	0.260	0.584	0.004	0.666	0.791	8	345.5	180–439
I80.81 Thrombophlebitis, superficial	0.906	0.856	0.957	0.057	0.678	0.810	8	82	65–115
B25.88 Other cytomegaloviral disease	0.819	0.688	0.963	0.044	0.800	0.812	7	266	161.5–384
I80.88 Thrombophlebitis, other site	0.651	0.478	0.817	0.007	0.661	0.670	7	328	237.5–541.5
C78.1 Sec. malig. neoplasm, mediastinum	0.843	0.805	0.881	0.011	0.711	0.541	7	397	246–491

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
M89.59	0.765	0.588	0.885	0.010	0.822	0.769	6	246.5	49.2–461.8
Osteolysis, unspecified site									
A41.51	0.638	0.279	0.930	0.015	0.485	0.565	6	584	466.2–644
Sepsis, other Gram-negative									
C79.9	0.727	0.546	0.888	0.009	0.596	0.410	6	358	119.8–477.8
Sec. malig. neoplasm, unspec. site									
C79.6	0.705	0.519	0.884	0.008	0.642	0.739	6	252	111.8–525
Sec. malig. neoplasm, ovary									
F32.0	0.778	0.680	0.870	0.007	0.348	0.263	6	72.5	21–477.2
Mild depressive episode									
B02.2	0.862	0.810	0.906	0.010	0.624	0.592	6	239	116.2–403.8
Zoster with neurological involv.									
N18.82	0.930	0.882	0.964	0.023	0.888	0.964	6	231.5	114.2–514.5
CKD with mineral/bone disorder									
A09	0.881	0.792	0.977	0.016	0.859	0.802	5	54	49–293
Gastroenteritis (any cause)									
I20.8	0.797	0.517	0.998	0.143	0.614	0.472	5	43	15–47
Other forms of angina pectoris									
N18.89	0.895	0.865	0.920	0.010	0.682	0.933	5	127	84–162
Other chronic kidney disease									
F33.2	0.343	0.211	0.471	0.002	0.710	0.821	5	161	57–189
Recurrent depressive disorder, severe									
J96.0	0.949	0.913	0.975	0.024	0.754	0.597	5	23	6–25
Acute respiratory failure									
I80.3	0.834	0.624	0.955	0.018	0.723	0.616	5	277	35–440
Phlebothrombosis, lower extremity									
I48.2	0.921	0.795	0.991	0.041	0.620	0.734	5	225	28–260
Chronic atrial fibrillation									
I25.13	0.805	0.532	0.988	0.033	0.709	0.755	5	23	6–66
Atherosclerotic HD, bypass graft									
S72.00	0.665	0.388	0.922	0.007	0.423	0.215	5	135	97–149
Fracture, femoral neck									
I48.11	0.821	0.671	0.968	0.011	0.812	0.802	5	402	201–496
Atrial flutter, type I									
C79.7	0.554	0.506	0.592	0.003	0.386	0.257	5	471	240–527
Sec. malig. neoplasm, adrenal gland									
E05.0	0.833	0.688	0.975	0.007	0.776	0.753	4	102	78–185
Thyrotoxicosis, Graves' disease									
I80.2	0.933	0.845	0.972	0.014	0.806	0.859	4	182.5	110.2–268
Phlebothrombosis, unspecified									
M81.89	0.710	0.492	0.926	0.005	0.442	0.574	4	128	27.2–250
Other osteoporosis w/o fracture									
C78.4	0.956	0.929	0.983	0.020	0.689	0.755	4	101.5	37.8–298.2
Sec. malig. neoplasm, small intestine									
B95.90	0.864	0.819	0.910	0.006	0.215	0.681	4	345.5	174.2–520
Streptococcus, unspecified									
C73	0.776	0.541	0.929	0.005	0.645	0.662	3	471	247–539
Malig. neoplasm, thyroid gland									
N17.92	0.902	0.718	0.998	0.051	0.805	0.814	3	18	14.5–26.5
Acute kidney failure, stage 2									
C79.82	0.930	0.799	0.996	0.031	0.763	0.844	3	3	3–4
Sec. malig. neoplasm, genital									
I42.0	0.358	0.139	0.679	0.102	0.484	0.421	3	192	117–318.5
Dilated cardiomyopathy									

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
J96.9 Respiratory failure, unspecified	0.888	0.732	0.986	0.013	0.765	0.696	3	120	99–126
I10 Essential hypertension	0.917	0.862	0.967	0.004	0.786	0.825	2	40	32–48
S42.20 Fracture, prox. humerus, unspecified	0.800	0.774	0.827	0.002	0.229	0.328	2	129	123.5–134.5
N18.0 End-stage renal disease	0.991	0.976	1.000	0.527	0.983	0.826	2	141.5	102.2–180.8
C77.4 Sec. malig. neoplasm, inguinal LN	0.774	0.699	0.851	0.003	0.922	0.503	2	14.5	7.8–21.2
I11.90 Hypertensive heart dis., unspec.	0.730	0.485	0.968	0.006	0.558	0.697	2	432	345.5–518.5
I48.19 Atrial flutter, other	0.885	0.760	0.998	0.051	0.915	0.825	2	299.5	248.2–350.8
B27.0 Infectious mononucleosis, EBV	0.999	0.996	1.000	0.287	0.860	0.962	2	53.5	34.8–72.2
S82.82 Fracture, lower leg, other parts	0.518	0.467	0.568	0.202	0.627	0.128	2	477	440.5–513.5
C77.0 Sec. malig. neoplasm, head/neck LN	0.941	0.884	0.993	0.022	0.923	0.832	2	21.5	12.8–30.2
D46.7 Other myelodysplastic syndromes	0.993	0.988	0.997	0.056	0.554	0.504	1	58	58–58
I25.12 Atherosclerotic HD, autol. graft	0.989	0.984	0.994	0.038	0.818	0.605	1	13	13–13
S42.21 Fracture, surgical neck of humerus	0.885	0.869	0.899	0.000	0.862	0.144	1	143	143–143
B00.9 Herpesviral infection, unspecified	0.798	0.779	0.818	0.000	0.530	0.579	1	99	99–99
B18.1 Chronic hepatitis B w/o delta	–	–	–	–	–	–	0	–	–
B18.2 Chronic hepatitis C	–	–	–	–	–	–	0	–	–
B24 HIV disease, unspecified	–	–	–	–	–	–	0	–	–
B35.1 Tinea unguium	–	–	–	–	–	–	0	–	–
B47.9 Mycetoma, unspecified	–	–	–	–	–	–	0	–	–
C79.4 Sec. malig. neoplasm, nervous system	–	–	–	–	–	–	0	–	–
D46.2 Refractory anaemia w/ excess blasts	–	–	–	–	–	–	0	–	–
E11.20 Type 2 diabetes w/ renal complic.	–	–	–	–	–	–	0	–	–
E13.90 Other specified diabetes mellitus	–	–	–	–	–	–	0	–	–
F32.8 Other depressive episodes	–	–	–	–	–	–	0	–	–
I80.8 Phlebothrombosis, other sites	–	–	–	–	–	–	0	–	–
S42.3 Fracture, humerus shaft	–	–	–	–	–	–	0	–	–
S52.50 Fracture, distal radius, unspec.	–	–	–	–	–	–	0	–	–

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
S52.51	–	–	–	–	–	–	0	–	–
Fracture, distal radius, extension									
S52.59	–	–	–	–	–	–	0	–	–
Fracture, distal radius, other									

Table 7. Predictive performance and prediction horizon across grouped diagnostic endpoints. Performance is summarised at two levels of aggregation: the bold rows give pooled metrics over all endpoints in a clinical category, the indented rows show the corresponding three-character ICD-10 sub-categories. Transformer AUROC is reported with 95% confidence intervals (CI) from repeated resampling; XGBoost and logistic regression AUROC are reported as medians; median average precision (AP) is additionally reported for the Transformer. Prediction horizon is the median time in days from the end of the laboratory input window to the diagnosis event, with interquartile range (IQR). Held-out test set comprising 1,579 sequences; *N* is the number of positive cases in the test set.

Endpoint	Transformer				XGBoost	Log. Reg.	N	Horizon	
	AUROC			AP	AUROC	AUROC		Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
Cardiovascular diseases	0.682	0.651	0.709	0.480	0.589	0.629	470	113	28.2–356
I10	0.917	0.862	0.967	0.004	0.786	0.825	263	92	25–279.5
Essential hypertension									
I48	0.741	0.695	0.788	0.188	0.671	0.752	110	187	52.2–411.2
Atrial fibrillation/flutter									
I80	0.601	0.544	0.654	0.088	0.582	0.638	99	177	72.5–360
Phlebitis/thrombophlebitis									
I50	0.683	0.616	0.749	0.156	0.594	0.631	76	175	70–482.5
Heart failure									
I25	0.705	0.640	0.769	0.108	0.561	0.673	70	147	19.2–362.8
Chronic ischaemic heart disease									
I11	0.815	0.700	0.902	0.119	0.625	0.727	22	191.5	37.8–547
Hypertensive heart disease									
I12	0.788	0.670	0.886	0.030	0.685	0.771	12	445	279.5–547.5
Hypertensive chronic kidney disease									
I20	0.755	0.558	0.896	0.061	0.726	0.632	11	47	15.5–90.5
Angina pectoris									
I21	0.714	0.615	0.807	0.011	0.527	0.579	11	243	66.5–448.5
Acute myocardial infarction									
I15	0.674	0.438	0.912	0.080	0.932	0.917	8	536.5	344.2–608
Secondary hypertension									
I82	0.386	0.099	0.662	0.007	0.520	0.475	8	205	111–317.5
Other venous embolism/thrombosis									
Bacterial infections	0.662	0.628	0.694	0.403	0.632	0.652	418	154	55.2–371.5
B96	0.660	0.618	0.703	0.255	0.623	0.628	219	211	71–443
Other bacterial agents (Gram-negative)									
B95	0.644	0.596	0.690	0.160	0.599	0.652	159	187	77.5–399
Streptococcus/Staphylococcus									
A41	0.735	0.667	0.794	0.166	0.635	0.727	84	233.5	77.2–480
Other sepsis									
A04	0.721	0.657	0.777	0.109	0.593	0.632	80	179.5	85–398.2
Other bacterial intestinal infections									
A09	0.881	0.792	0.977	0.016	0.859	0.802	79	213	77–432
Other gastroenteritis/colitis									
B99	0.588	0.499	0.669	0.063	0.542	0.600	61	160	76–438
Other/unspec. infectious disease									
Metastatic disease	0.725	0.695	0.758	0.483	0.715	0.729	366	165	44.2–338.8
C78	0.818	0.791	0.841	0.403	0.793	0.804	218	149.5	33.5–316
Sec. malig. neoplasm, resp./digestive									
C79	0.613	0.565	0.663	0.127	0.538	0.570	133	219	85–398
Sec. malig. neoplasm, other sites									

Continued on next page

Endpoint	Transformer				XGBoost	Log. Reg.		Horizon	
	AUROC			AP	AUROC	AUROC	N	Median	IQR
	Median	CI ₉₅ Low	CI ₉₅ High	Median	Median	Median		(days)	(days)
C77 Sec. malig. neoplasm, lymph nodes	0.777	0.740	0.813	0.215	0.735	0.765	123	147	35–318
Kidney disease	0.741	0.705	0.774	0.372	0.720	0.742	243	232	73.5–396
N17 Acute kidney failure	0.661	0.614	0.702	0.156	0.620	0.691	141	235	70–386
N18 Chronic kidney disease	0.887	0.854	0.914	0.458	0.855	0.866	139	222	84.5–431.5
Z99 Dependence on renal dialysis	0.756	0.661	0.844	0.126	0.850	0.784	36	387.5	164–521.2
Thyroid disorders	0.580	0.538	0.623	0.158	0.593	0.592	196	175	41–364
E03 Hypothyroidism	0.667	0.620	0.712	0.138	0.621	0.608	127	127	23.5–331
E04 Nontoxic goitre	0.580	0.512	0.645	0.034	0.570	0.602	48	268.5	106.2–490.8
E05 Thyrotoxicosis	0.677	0.521	0.814	0.023	0.716	0.718	16	111	65–448
E06 Thyroiditis	0.550	0.371	0.728	0.021	0.638	0.655	13	237	75–364
All-cause mortality	0.697	0.653	0.737	0.213	0.656	0.660	187	346	195–531.5
Osteolyses	0.833	0.800	0.865	0.258	0.825	0.850	102	189.5	70.2–350.8
Type 2 Diabetes	0.646	0.574	0.709	0.074	0.571	0.614	65	189	70–371
Fungal infections	0.752	0.691	0.814	0.101	0.664	0.500	52	151	69.8–266
B37 Candidiasis	0.754	0.691	0.818	0.095	0.675	0.729	47	153	73.5–276.5
B44 Aspergillosis	0.854	0.750	0.943	0.052	0.858	0.841	11	196	59.5–318
Depressive disorders	0.575	0.496	0.655	0.049	0.511	0.494	47	167	56.5–383
GI complications	0.687	0.603	0.766	0.046	0.520	0.595	38	215.5	141–386.8
Viral infections	0.743	0.643	0.835	0.080	0.628	0.512	26	261	95.5–437.8
B02 Herpes zoster	0.727	0.611	0.832	0.029	0.488	0.324	17	291	178–480
B25 Cytomegaloviral disease	0.819	0.688	0.963	0.044	0.800	0.812	7	266	161.5–384
B27 Infectious mononucleosis	0.999	0.996	1.000	0.287	0.860	0.962	2	53.5	34.8–72.2
MDS	0.711	0.580	0.830	0.081	0.576	0.603	21	358	318–452

Table 8. External validation performance per ICD-10-GM endpoint on MIMIC-IV and MMRF CoMMpass. For each of the 162 target endpoints we report area under the receiver-operating-characteristic curve (AUROC) with bootstrap 95% confidence interval, average precision (AP), and the number of positive samples (N) on the full MIMIC-IV and the MMRF CoMMpass IA24 datasets without finetuning. AUROC and AP values are bootstrap medians. N is computed as $\lfloor \pi \times N_{\text{eval}} \rfloor$ where π is the median positive-class prevalence and $N_{\text{eval}}^{\text{MIMIC}} = 8150$, $N_{\text{eval}}^{\text{MMRF}} = 7347$ samples. Endpoints with zero positive cases in a given cohort are reported as “—”. Rows are ordered by the larger of the two cohort prevalences, descending.

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
All-cause mortality	0.635	0.622	0.648	0.432	2356	0.691	0.677	0.704	0.324	1582
All-cause mortality										
M89.59	—	—	—	—	0	0.553	0.514	0.587	0.227	1378
Osteolysis, unspec. site										
M89.55	—	—	—	—	0	0.595	0.560	0.631	0.243	1378
Osteolysis, femur										

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
M89.58 Osteolysis, other site	0.380	0.229	0.586	0.001	8	0.571	0.534	0.606	0.237	1378
M89.50 Osteolysis, multi-site/unspec.	–	–	–	–	0	0.582	0.546	0.617	0.231	1378
I10.91 Essential HTN, with crisis	0.516	0.498	0.536	0.140	1068	0.741	0.700	0.779	0.102	254
I10.01 Malignant essential HTN	0.528	0.510	0.549	0.147	1068	0.675	0.631	0.721	0.063	254
I10 Essential primary hypertension	0.472	0.453	0.492	0.138	1068	0.551	0.502	0.597	0.038	254
I10.90 Essential HTN, unspec., no crisis	0.533	0.515	0.552	0.144	1068	0.692	0.647	0.736	0.074	254
I10.00 Essential hypertension, benign	0.534	0.517	0.553	0.145	1068	0.705	0.663	0.749	0.088	254
N17.92 AKI, KDIGO stage 2	0.642	0.623	0.659	0.166	844	0.669	0.617	0.711	0.028	126
N17.93 AKI, stage 3	0.627	0.607	0.646	0.162	844	0.685	0.640	0.722	0.025	126
N17.9 AKI, unspecified	0.635	0.617	0.655	0.165	844	0.665	0.613	0.712	0.035	126
N17.91 AKI, stage 1	0.605	0.585	0.624	0.157	844	0.652	0.600	0.696	0.025	126
N17.99 AKI, stage unspec.	0.606	0.585	0.624	0.155	844	0.644	0.588	0.688	0.024	126
I12.00 Hypertensive renal disease, w/insuff.	0.839	0.822	0.857	0.180	370	0.716	0.681	0.746	0.255	761
E11.20 T2DM with kidney complications	0.797	0.769	0.825	0.060	139	0.740	0.709	0.765	0.209	566
F32.9 Depressive episode, unspec.	0.495	0.469	0.519	0.071	541	0.558	0.416	0.694	0.002	15
N18.89 Other CKD, unspec. stage	0.735	0.712	0.756	0.095	353	0.739	0.708	0.767	0.188	536
N18.82 Other CKD (non-standard)	0.648	0.621	0.674	0.071	353	0.727	0.700	0.752	0.159	536
N18.2 CKD stage 2	0.686	0.603	0.766	0.008	29	0.713	0.681	0.741	0.152	536
N18.0 End-stage renal disease	0.820	0.780	0.859	0.116	114	0.744	0.715	0.772	0.193	536
N18.9 CKD, unspecified	0.794	0.773	0.813	0.124	353	0.756	0.723	0.781	0.198	536
N18.4 CKD stage 4	0.853	0.814	0.887	0.119	102	0.750	0.719	0.776	0.205	536
N18.3 CKD stage 3	0.846	0.829	0.864	0.102	218	0.753	0.722	0.780	0.212	536
N18.5 CKD stage 5	0.830	0.787	0.869	0.114	114	0.751	0.724	0.777	0.192	536
J96.0 Acute respiratory failure	0.477	0.444	0.505	0.050	355	0.539	0.429	0.644	0.012	53
J96.00 Acute respir. failure, type I (hypoxic)	0.552	0.521	0.582	0.061	350	0.601	0.512	0.682	0.014	53
I50.14 Left heart failure, NYHA IV	0.600	0.567	0.630	0.068	341	0.597	0.514	0.663	0.013	61
I50.12 Left heart failure, NYHA II	0.688	0.664	0.714	0.086	341	0.550	0.457	0.634	0.011	61

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
I50.13 Left heart failure, NYHA III	0.689	0.666	0.714	0.081	341	0.604	0.509	0.688	0.027	61
C78.6 Sec. malig., peritoneum	0.755	0.733	0.775	0.092	332	–	–	–	–	–
I11.00 Hypertensive HD, with HF	0.629	0.592	0.665	0.035	173	0.616	0.573	0.655	0.054	290
I25.19 Atherosclerotic HD, unspec.	0.672	0.647	0.700	0.061	282	0.646	0.577	0.723	0.007	38
I25.12 Atherosclerotic HD, 2-vessel	0.707	0.681	0.736	0.071	282	0.696	0.625	0.765	0.008	38
I25.11 Atherosclerotic HD, native coronary	0.690	0.664	0.717	0.065	282	0.614	0.536	0.684	0.009	54
I25.13 Atherosclerotic HD, 3-vessel	0.699	0.674	0.726	0.068	282	0.691	0.629	0.755	0.007	38
I48.9 AF/atrial flutter, unspec.	0.658	0.629	0.686	0.060	273	0.639	0.570	0.702	0.019	88
E11.91 T2DM w/o compl. (decompensated)	0.592	0.555	0.623	0.065	265	0.780	0.666	0.871	0.009	23
E11.90 T2DM w/o complications	0.586	0.551	0.620	0.049	265	0.632	0.527	0.732	0.005	23
E03.9 Hypothyroidism, unspec.	0.486	0.453	0.519	0.033	263	0.647	0.541	0.742	0.006	21
I48.2 Atrial fibrillation, chronic	0.668	0.634	0.699	0.068	258	0.660	0.591	0.715	0.019	88
I15.10 Renovascular hypertension	–	–	–	–	0	0.696	0.653	0.739	0.066	254
I11.90 Hypertensive HD, no HF, no crisis	0.452	0.441	0.464	0.000	1	0.702	0.656	0.746	0.090	254
C78.7 Sec. malig., liver	0.777	0.749	0.802	0.111	232	–	–	–	–	–
A41.9 Sepsis, unspecified	0.549	0.513	0.588	0.035	212	0.622	0.575	0.668	0.031	159
C79.88 Sec. malig., other sites	0.749	0.725	0.775	0.060	208	–	–	–	–	–
B37.0 Candidal stomatitis	0.607	0.570	0.642	0.035	189	0.480	0.149	0.885	0.003	7
B96.2 Escherichia coli	0.596	0.559	0.635	0.027	163	0.576	0.511	0.636	0.010	65
I48.0 Atrial fibrillation, paroxysmal	0.639	0.600	0.676	0.035	162	0.682	0.616	0.737	0.021	88
A04.7 Enterocolitis, C. difficile	0.518	0.473	0.558	0.025	161	0.511	0.395	0.619	0.007	31
A04.70 Enterocolitis, C. diff., w/o mega-colon	0.598	0.557	0.637	0.028	161	0.554	0.408	0.675	0.006	31
A41.1 Sepsis, other Staphylococcus	0.712	0.566	0.834	0.005	14	0.631	0.576	0.680	0.040	159
A41.58 Sepsis, other Gram-negative	0.735	0.605	0.855	0.003	8	0.644	0.596	0.689	0.042	159
A41.51 Sepsis, E. coli	0.688	0.625	0.744	0.013	51	0.608	0.558	0.656	0.034	159
I25.22 Old MI, ≥1 year	0.726	0.689	0.766	0.045	147	0.671	0.610	0.737	0.007	38
C78.5 Sec. malig., large intestine	0.737	0.699	0.769	0.044	145	–	–	–	–	–

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
I50.01	0.580	0.530	0.626	0.025	130	0.673	0.596	0.737	0.019	61
Right heart failure (secondary)										
C79.5	0.560	0.512	0.605	0.020	128	–	–	–	–	–
Sec. malig., bone/marrow										
C77.2	0.748	0.717	0.776	0.033	126	–	–	–	–	–
Sec. malig., intra-abd. LN										
N17.8	0.589	0.490	0.682	0.005	25	0.670	0.622	0.721	0.031	126
Other acute kidney failure										
I50.9	0.611	0.567	0.656	0.024	123	0.608	0.521	0.686	0.025	61
Heart failure, unspec.										
C79.82	0.746	0.711	0.780	0.033	113	–	–	–	–	–
Sec. malig., genital organs										
B95.2	0.635	0.592	0.676	0.022	111	0.587	0.519	0.654	0.010	65
Strep. grp. D/Enterococcus										
I80.1	0.499	0.442	0.565	0.015	103	0.606	0.531	0.688	0.016	52
Phlebitis/throm., femoral vein										
C78.0	0.774	0.734	0.816	0.037	92	–	–	–	–	–
Sec. malig., lung										
K92.2	0.691	0.614	0.768	0.014	39	0.534	0.467	0.590	0.012	90
GI haemorrhage, unspec.										
I48.1	0.769	0.717	0.814	0.023	55	0.661	0.596	0.716	0.018	88
Atrial fibrillation, persistent										
I48.19	0.799	0.748	0.844	0.024	55	0.635	0.566	0.692	0.023	88
Persistent AF, other (non-std)										
I48.11	0.776	0.721	0.826	0.023	55	0.655	0.574	0.717	0.036	88
Persistent AF, longstanding (non-std)										
I48.10	0.767	0.713	0.818	0.021	55	0.690	0.626	0.739	0.036	88
Persistent AF (non-standard)										
I80.2	0.562	0.502	0.620	0.016	87	0.433	0.338	0.526	0.011	52
Phlebitis, other deep LE veins										
I21.4	0.633	0.585	0.681	0.019	87	0.742	0.676	0.803	0.008	31
Acute subendocardial MI (NSTEMI)										
B96.8	0.499	0.438	0.564	0.011	81	0.581	0.506	0.650	0.011	65
Other specified bacterial agents										
B95.90	0.557	0.506	0.608	0.013	81	0.530	0.454	0.605	0.007	55
Other gram-positive aerobes										
R73.9	0.525	0.463	0.587	0.012	77	0.856	0.618	0.980	0.015	8
Hyperglycaemia, unspecified										
Z99.2	0.826	0.769	0.880	0.122	76	0.899	0.824	0.967	0.072	16
Dependence on renal dialysis										
C78.4	0.726	0.686	0.762	0.018	74	–	–	–	–	–
Sec. malig., small intestine										
C79.9	0.573	0.506	0.641	0.012	69	–	–	–	–	–
Sec. malig., unspec. site										
C77.5	0.700	0.650	0.745	0.015	69	–	–	–	–	–
Sec. malig., intrapelvic LN										
B96.6	0.936	0.931	0.942	0.500	1	0.567	0.498	0.635	0.010	65
Bacteroides fragilis										
B96.5	0.676	0.599	0.751	0.007	30	0.575	0.499	0.643	0.010	65
Pseudomonas										
J96.9	0.536	0.466	0.602	0.010	64	0.590	0.483	0.687	0.011	53
Respiratory failure, unspec.										
C78.8	0.639	0.592	0.684	0.012	64	–	–	–	–	–
Sec. malig., other digestive										

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
B95.7 Other staphylococci	0.549	0.481	0.619	0.011	63	0.469	0.401	0.537	0.006	55
B95.6 Staphylococcus aureus	0.575	0.515	0.638	0.010	60	0.474	0.408	0.537	0.006	55
C79.3 Sec. malig., brain	0.682	0.610	0.748	0.016	56	–	–	–	–	–
B95.48 Other specified streptococci	0.428	0.327	0.534	0.002	15	0.425	0.341	0.509	0.007	55
B95.8 Unspecified strepto- cocci/staphylococci	0.143	0.081	0.208	0.001	7	0.553	0.465	0.641	0.009	55
J96.09 Acute respir. failure, unspec. type	0.378	0.060	0.818	0.100	3	0.626	0.551	0.698	0.011	53
B99 Other infectious disease, unspec.	0.571	0.453	0.692	0.003	15	0.732	0.653	0.796	0.023	53
I80.81 Phlebitis/throm., deep UE veins	0.585	0.508	0.660	0.009	49	0.681	0.598	0.756	0.013	52
I80.80 Phlebitis, superficial UE veins	0.595	0.487	0.699	0.012	31	0.702	0.607	0.788	0.024	52
I80.8 Phlebitis/throm., other localiz.	0.331	0.262	0.407	0.003	31	0.411	0.318	0.499	0.008	52
I80.88 Phlebitis/throm., other localiz.	0.467	0.357	0.563	0.005	31	0.633	0.554	0.705	0.011	52
I80.3 Phlebitis/throm., LE incl. DVT NOS	0.511	0.435	0.591	0.005	38	0.590	0.514	0.669	0.014	52
I80.28 Phlebitis, other LE veins	0.324	0.142	0.509	0.001	6	0.537	0.456	0.604	0.022	52
B02.2 Zoster with neurological involv.	0.399	0.308	0.497	0.002	20	0.694	0.608	0.777	0.020	52
B02.9 Herpes zoster, uncomplicated	0.774	0.602	0.906	0.004	8	0.643	0.576	0.702	0.010	52
C78.2 Sec. malig., pleura	0.715	0.652	0.778	0.015	51	–	–	–	–	–
I82.9 Embolism/thrombosis, unspec. veins	–	–	–	–	0	0.547	0.447	0.638	0.013	50
Z99.1 Respirator dependence, long-term	–	–	–	–	0	0.594	0.500	0.677	0.009	49
B18.1 Chronic viral hepatitis B	0.527	0.450	0.611	0.007	45	0.674	0.540	0.788	0.003	12
S72.10 Petrochanteric fracture	0.759	0.641	0.861	0.007	18	0.425	0.339	0.516	0.005	44
S72.00 Femoral neck fracture, unspec.	0.707	0.619	0.786	0.008	29	0.448	0.349	0.533	0.005	44
S82.6 Fracture, lateral malleolus	–	–	–	–	0	0.524	0.454	0.594	0.005	41
S82.82 Fracture, other lower leg	0.381	0.234	0.548	0.001	10	0.521	0.431	0.603	0.005	41
C77.8 Sec. malig., LN multiple	0.802	0.760	0.840	0.015	40	–	–	–	–	–
I25.9 Chronic ischaemic HD, unspec.	–	–	–	–	0	0.691	0.615	0.767	0.008	38
C77.0 Sec. malig., head/neck LN	0.548	0.444	0.643	0.006	35	–	–	–	–	–
B37.88 Candidiasis, other sites	0.559	0.468	0.645	0.005	31	0.813	0.615	0.936	0.003	7

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
A09.0	0.577	0.484	0.657	0.005	31	0.491	0.385	0.599	0.003	27
Infectious gastroenteritis										
A09	0.411	0.325	0.498	0.018	31	0.696	0.576	0.795	0.007	27
Diarrhoea/gastroent., infectious										
A09.9	0.494	0.393	0.595	0.004	31	0.511	0.400	0.630	0.004	27
Gastroenteritis, unspec.										
C77.9	0.578	0.498	0.658	0.009	31	–	–	–	–	–
Sec. malig., LN unspec.										
C79.6	0.839	0.794	0.882	0.013	29	–	–	–	–	–
Sec. malig., ovary										
C77.1	0.681	0.617	0.744	0.006	28	–	–	–	–	–
Sec. malig., thoracic LN										
D46.9	0.749	0.658	0.826	0.014	28	–	–	–	–	–
Myelodysplastic syndrome, unspec.										
M80.88	0.747	0.591	0.873	0.007	13	0.319	0.192	0.453	0.003	26
Osteoporosis with pathol. fracture										
C79.2	0.657	0.573	0.736	0.005	23	–	–	–	–	–
Sec. malig., skin										
E13.90	0.078	0.073	0.085	0.000	1	0.714	0.583	0.835	0.008	23
Other specified DM, no compl.										
S42.20	0.554	0.331	0.740	0.051	5	0.639	0.510	0.772	0.006	21
Fracture, prox. humerus, unspec.										
S42.3	0.210	0.006	0.421	0.000	2	0.680	0.566	0.775	0.005	21
Fracture, humerus shaft										
S42.21	0.512	0.175	0.898	0.143	4	0.600	0.502	0.719	0.004	21
Fracture, surg. neck of humerus										
E03.8	0.474	0.353	0.595	0.003	15	0.524	0.368	0.661	0.021	21
Other hypothyroidism										
I20.8	0.718	0.598	0.835	0.334	2	0.423	0.254	0.584	0.003	20
Other forms of angina pectoris										
I20.0	0.832	0.470	0.956	0.003	4	0.463	0.314	0.606	0.003	20
Unstable angina										
C77.4	0.595	0.483	0.689	0.004	20	–	–	–	–	–
Sec. malig., inguinal/LE LN										
B00.9	0.527	0.439	0.603	0.003	19	0.517	0.505	0.528	0.500	1
Herpesviral infection, unspec.										
B18.2	0.490	0.387	0.596	0.003	18	0.615	0.467	0.721	0.002	12
Chronic viral hepatitis C										
B25.88	0.642	0.514	0.751	0.004	18	–	–	–	–	–
Other CMV disease										
I42.0	0.711	0.592	0.808	0.005	15	0.640	0.361	0.835	0.002	7
Dilated cardiomyopathy										
F33.2	0.285	0.158	0.454	0.037	8	0.656	0.442	0.819	0.009	15
Recurrent depression, severe										
F32.2	–	–	–	–	0	0.638	0.495	0.790	0.004	15
Severe depression w/o psychosis										
F32.0	–	–	–	–	0	0.463	0.352	0.589	0.003	15
Mild depressive episode										
F32.8	–	–	–	–	0	0.521	0.339	0.704	0.003	15
Other depressive episodes										
F33.1	0.728	0.585	0.841	0.126	3	0.642	0.438	0.805	0.004	15
Recurrent depression, moderate										
F32.1	–	–	–	–	0	0.617	0.434	0.775	0.003	15
Moderate depressive episode										
S52.51	0.670	0.512	0.814	0.072	6	0.520	0.371	0.659	0.004	15
Distal radius fracture, Colles										

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
S52.59 Other distal radius fractures	0.563	0.384	0.832	0.126	4	0.654	0.520	0.791	0.003	15
S52.50 Distal radius fracture, unspec.	0.810	0.801	0.820	0.001	2	0.639	0.528	0.752	0.004	15
E05.0 Hyperthyroidism, Graves'	0.413	0.227	0.567	0.002	14	0.709	0.699	0.717	0.500	1
C73 Malig. neoplasm, thyroid gland	0.590	0.459	0.703	0.003	14	0.674	0.462	0.888	0.167	3
S32.01 Fracture, L1 vertebra	0.587	0.474	0.706	0.002	11	0.314	0.143	0.492	0.001	10
B44.1 Pulmonary aspergillosis	0.496	0.392	0.611	0.002	11	–	–	–	–	–
D46.7 Other myelodysplastic syndromes	0.345	0.272	0.419	0.040	9	–	–	–	–	–
C78.1 Sec. malig. neoplasm, mediastinum	0.581	0.414	0.744	0.002	8	–	–	–	–	–
B27.0 Infectious mononucleosis, EBV	0.769	0.649	0.881	0.003	8	0.467	0.126	0.809	0.250	2
E05.9 Hyperthyroidism, unspec.	0.716	0.465	0.903	0.004	7	0.256	0.245	0.266	0.000	1
B35.1 Tinea unguium	0.375	0.116	0.627	0.001	7	0.716	0.706	0.726	0.500	1
C79.4 Sec. malig., other nerv. system	0.769	0.640	0.878	0.002	7	–	–	–	–	–
E06.3 Autoimmune thyroiditis (Hashimoto)	0.699	0.490	0.885	0.003	6	0.501	0.470	0.529	0.167	2
M81.89 Other osteoporosis w/o fracture	0.402	0.285	0.545	0.001	5	0.412	0.016	0.812	0.167	2
C79.7 Sec. malig., adrenal gland	0.833	0.743	0.925	0.003	5	–	–	–	–	–
C77.3 Sec. malig., axillary/UE LN	0.847	0.754	0.941	0.003	5	–	–	–	–	–
E04.2 Nontoxic multinodular goitre	0.243	0.004	0.707	0.100	3	0.681	0.519	0.906	0.126	4
E04.9 Nontoxic goitre, unspec.	–	–	–	–	0	0.694	0.548	0.906	0.126	4
D46.2 RAEB	0.913	0.820	0.992	0.009	4	–	–	–	–	–
B24 HIV disease, unspecified	0.257	0.109	0.357	0.000	3	–	–	–	–	–
M81.99 Osteoporosis, unspecified	–	–	–	–	0	0.827	0.678	0.971	0.002	2
B47.9 Mycetoma, unspecified	–	–	–	–	0	–	–	–	–	–
R99 III-defined causes of mortality	–	–	–	–	0	–	–	–	–	–

Table 9. External validation performance across grouped diagnostic endpoints on MIMIC-IV and MMRF CoMMpass. Performance is summarised at two levels of aggregation: the bold rows give pooled metrics over all endpoints in a clinical category, the indented rows show the corresponding three-character ICD-10 sub-categories. AUROC and average precision (AP) are bootstrap medians; CI₉₅ bounds are the 2.5th and 97.5th percentiles. The number of positive samples (N) is computed as $\lfloor \pi \times N_{\text{eval}} \rfloor$ where π is the median positive-class prevalence and $N_{\text{eval}}^{\text{MIMIC}} = 8150$, $N_{\text{eval}}^{\text{MMRF}} = 7347$ samples. Endpoints with zero positive cases in a given cohort are reported as “—”. Rows are ordered by the larger of the two cohort prevalences, descending.

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
All-cause mortality	0.635	0.622	0.648	0.432	2356	0.691	0.677	0.704	0.324	1582
Cardiovascular diseases	0.593	0.578	0.608	0.277	1645	0.669	0.648	0.690	0.141	674
I10	0.472	0.453	0.492	0.138	1068	0.551	0.502	0.597	0.038	254
Essential hypertension										
I48	0.643	0.620	0.666	0.091	411	0.682	0.616	0.737	0.021	88
Atrial fibrillation/flutter										
I80	0.492	0.454	0.531	0.021	161	0.549	0.465	0.616	0.009	52
Phlebitis/thrombophlebitis										
I50	0.652	0.629	0.675	0.090	423	0.590	0.497	0.673	0.025	61
Heart failure										
I25	0.686	0.664	0.713	0.074	321	0.591	0.501	0.670	0.008	54
Chronic ischaemic heart disease										
I11	0.621	0.585	0.657	0.033	169	0.619	0.577	0.659	0.054	289
Hypertensive heart disease										
I21	0.633	0.585	0.681	0.019	87	0.742	0.676	0.803	0.008	31
Acute myocardial infarction										
I82	—	—	—	—	—	0.547	0.447	0.638	0.013	50
Other venous embolism/thrombosis										
Osteolyses	0.279	0.131	0.467	0.001	8	0.579	0.542	0.614	0.228	1378
Kidney disease	0.694	0.677	0.709	0.259	1081	0.668	0.638	0.693	0.099	501
N17	0.591	0.572	0.609	0.125	748	0.683	0.631	0.724	0.027	126
Acute kidney failure										
N18	0.827	0.812	0.840	0.257	634	0.753	0.722	0.779	0.209	536
Chronic kidney disease										
Z99	0.824	0.767	0.876	0.093	75	0.653	0.572	0.722	0.030	64
Dependence on renal dialysis										
Metastatic disease	0.667	0.650	0.685	0.191	844	—	—	—	—	—
C78	0.763	0.747	0.778	0.173	582	—	—	—	—	—
Sec. malig. neoplasm, resp./digestive										
C79	0.595	0.573	0.617	0.083	479	—	—	—	—	—
Sec. malig. neoplasm, other sites										
C77	0.725	0.698	0.750	0.064	249	—	—	—	—	—
Sec. malig. neoplasm, lymph nodes										
Bacterial infections	0.562	0.542	0.582	0.119	746	0.603	0.571	0.632	0.063	342
B96	0.574	0.543	0.607	0.043	261	0.570	0.506	0.629	0.009	65
Other bacterial agents (Gram-negative)										
B95	0.555	0.524	0.586	0.046	291	0.551	0.499	0.597	0.014	104
Streptococcus/Staphylococcus										
A41	0.582	0.549	0.615	0.047	265	0.621	0.572	0.667	0.032	159
Other sepsis										
A04	0.580	0.539	0.621	0.028	161	0.616	0.498	0.721	0.007	31
Other bacterial intestinal infections										
A09	0.411	0.325	0.498	0.018	31	0.696	0.576	0.795	0.007	27
Other gastroenteritis/colitis										
J96	0.538	0.510	0.567	0.063	389	0.603	0.519	0.686	0.014	53
Respiratory failure										
B99	0.571	0.453	0.692	0.003	15	0.732	0.653	0.796	0.023	53
Other/unspec. infectious disease										
Depressive disorders	0.489	0.464	0.515	0.066	499	0.550	0.390	0.686	0.002	15

Continued on next page

Endpoint	MIMIC-IV					MMRF CoMMpass				
	AUROC			AP	N	AUROC			AP	N
	Median	CI ₉₅ Low	CI ₉₅ High	Median		Median	CI ₉₅ Low	CI ₉₅ High	Median	
Type 2 Diabetes	0.650	0.623	0.675	0.081	363	0.682	0.653	0.706	0.099	436
Thyroid disorders	0.471	0.439	0.503	0.036	289	0.598	0.488	0.698	0.009	30
E03	0.478	0.445	0.510	0.032	258	0.536	0.394	0.674	0.010	21
Hypothyroidism										
E04	–	–	–	–	–	–	–	–	–	–
Nontoxic goitre										
E05	0.669	0.577	0.754	0.004	20	–	–	–	–	–
Thyrotoxicosis										
E06	0.699	0.490	0.885	0.003	6	–	–	–	–	–
Thyroiditis										
Fungal infections	0.563	0.529	0.598	0.035	215	0.615	0.387	0.860	0.005	8
B37	0.563	0.528	0.598	0.034	207	0.588	0.340	0.910	0.005	7
Candidiasis										
B44	0.496	0.392	0.611	0.002	11	–	–	–	–	–
Aspergillosis										
Viral infections	0.552	0.504	0.598	0.020	126	0.629	0.567	0.690	0.014	67
B02	0.560	0.456	0.657	0.004	26	0.656	0.587	0.722	0.011	52
Herpes zoster										
B25	0.642	0.514	0.751	0.004	18	–	–	–	–	–
Cytomegaloviral disease										
B27	0.769	0.649	0.881	0.003	8	0.467	0.126	0.809	0.250	2
Infectious mononucleosis										
GI complications	0.691	0.614	0.768	0.014	39	0.534	0.467	0.590	0.012	90
MDS	0.663	0.584	0.740	0.014	41	–	–	–	–	–

Supplementary Results

Information extraction from unstructured clinical documents

Extraction accuracy was evaluated by exact match against expert annotations across 60 clinically relevant variables grouped into 15 clinical domains (Extended Data Fig. 8d, Table 4). Accuracy exceeded 0.90 for most variables, including laboratory analyte values (accuracy > 0.95) and medication active substances, doses, and units (accuracy > 0.90). Core diagnostic fields, treatment line variables (0.85 to 0.95), refractory status, and cytogenetic abnormality fields also exceeded 0.90. Date fields showed greater variability, with strict exact-match accuracy ranging from 0.40 to 0.95 depending on domain. Under relaxed date matching permitting a one-month temporal window, accuracy improved by up to 45 percentage points for affected variables (Table 4).

External validation on 14 documents from four independent institutions confirmed transferability of extraction performance (Table 5). Demographic fields, diagnosis, bone marrow assessment, and death date achieved perfect or near-perfect accuracy. Medication extraction showed slightly lower accuracy on external documents (active substance 0.895, dose 0.870) compared with internal documents, consistent with greater variability in document formatting across institutions. Laboratory measurement accuracy remained above 0.85 across all external fields.

Comparison with non-sequential baselines

Per-endpoint and per-diagnosis performance comparisons between the transformer, XGBoost, and logistic regression are provided in Supplementary Tables 6, 7. Baseline methodology is described in Methods.

Model calibration

Model calibration was assessed using Platt scaling fitted independently per diagnosis group on the validation set. Predictions were binned into ten equally large quantiles (Figure 5b). After scaling, all three models were well calibrated, and their quantile bins overlapped closely and spanned a comparable range of predicted probabilities for each endpoint. The majority of predictions fell within the 0.0–0.2 range, where the binned predictions of all three models tracked the calibration diagonal closely. For the higher-prevalence endpoints (all-cause mortality, metastatic disease, cardiovascular disease, kidney disease and bacterial infections), predictions extended past 0.3 predicted probability, reaching approximately 0.5–0.6 for metastatic and cardiovascular disease, and remained close to the diagonal, with the highest-probability bins tending to fall slightly below it, indicating mild over-prediction in the sparsely populated upper range. For the low-prevalence endpoints (myelodysplastic syndromes, type 2 diabetes and fungal infections), predictions were confined to the lower-frequency region, limiting the assessable calibration range but showing close agreement with the diagonal within that region. Notably, the transformer’s discrimination advantage over the non-sequential baselines (Figure 5a) did not translate into a corresponding calibration difference: after per-group Platt scaling, the XGBoost and logistic regression bins coincided with those of the transformer at equivalent predicted probabilities across all eight diagnosis groups.

Subgroup stratification and ablation

The transformer was evaluated under stratified and ablated conditions (Figure 7); all values are median AUROC across all diagnoses. Performance was broadly preserved across age strata (Figure 7b): median AUROC 0.70–0.71 across the three older strata (55–64: 0.708; 65–74: 0.710; 75+: 0.703) and 0.762 in the youngest (18–54). Female patients achieved higher median AUROC than male patients (0.721 vs. 0.678), with overlapping 95% bootstrap confidence intervals; the gap likely reflects cohort composition, as the ovarian cancer subcohort contributing approximately half the training data is exclusively female.

Ablating demographic inputs or the imputation module had divergent effects on the internal and external cohorts (Figure 7a). On the internal test set, the full model (sex, age, and laboratory imputation) achieved a median AUROC of 0.740; removing sex (−0.004), age (+0.004), both demographic variables (+0.002), or laboratory imputation (+0.001) each changed median AUROC by less than 0.01, indicating that internal discrimination derives almost entirely from the laboratory measurements. Using the full model while disabling the ensemble reduced median AUROC by 0.037. On the external cohorts the same ablations produced larger degradations. On MIMIC-IV (full model 0.642), removing sex, age, or both lowered AUROC by 0.013 to 0.023, with the largest drop for removing both demographic variables (−0.023); omitting imputation reduced AUROC by 0.021. On MMRF CoMMpass (full model 0.616), removing sex produced the largest single drop (−0.029), followed by age (−0.024), both demographics (−0.019), and no imputation (−0.015). The contrast indicates that demographic features and the imputation module, while largely redundant with the laboratory signal internally, contribute materially to cross-cohort generalisation, consistent with the masking analysis: on MMRF the derived glomerular filtration estimates are unavailable, so the reconstruction path that internally compensates for missing demographic information is absent.

All values reported here are medians across individual ICD codes rather than the grouped endpoints used in the main text. At the level of individual codes, differences between datasets partly reflect inconsistent definitions and coding practices across cohorts rather than genuine differences in model performance. Grouping subcodes into broader categories reconciles these definitional discrepancies, so the cross-dataset differences are larger at the per-code level than for the grouped endpoints.

Per-diagnosis performance within grouped endpoints

Performance varied across individual ICD-10 codes within several grouped endpoints (Supplementary Table 6). Within the metastatic group, C78 (secondary malignant neoplasms of respiratory and digestive organs; AUROC 0.82, 2.9-fold enrichment) was the strongest sub-category, with peritoneal metastases (C78.6; AUROC 0.89, 5.1-fold enrichment) the strongest individual contributor, whereas other-site metastases (C79; AUROC 0.61, 1.5-fold) showed only modest enrichment. Kidney disease discrimination was driven primarily by chronic kidney disease (N18; AUROC 0.89, 5.2-fold enrichment) rather than acute kidney failure (N17; AUROC 0.66, 1.8-fold). End-stage renal disease (N18.5; AUROC 0.90) and dialysis dependence (Z99.2; AUROC 0.81) showed the highest enrichment of any code present in at least 1% of test cases (both approximately 25-fold), although their confidence intervals were wide. Sepsis (A41; AUROC 0.74, 3.1-fold enrichment) showed stronger enrichment than the more prevalent gram-positive (B95; AUROC 0.64, 1.6-fold) and gram-negative (B96; AUROC 0.66, 1.9-fold) organism codes that dominated the bacterial infections group. Among individual cardiovascular endpoints, atrial fibrillation (I48; AUROC 0.74, 2.8-fold) and chronic ischaemic heart disease (I25; AUROC 0.71, 2.5-fold) exceeded the group-level AUROC of 0.68, while heart failure (I50; AUROC 0.68, 3.3-fold) showed the strongest enrichment at comparable discrimination. Among rare diagnoses with fewer than 10 positive cases in the test set, confidence intervals widened substantially. Full per-diagnosis results including confidence intervals for all 162 endpoints are provided in Supplementary Table 6.

Supplementary Methods

Clinically defined target entities

The following clinical entities, identified by oncology and radiology experts, informed the design of extraction templates, validation rules, and the evaluation framework (Supplementary Table 4).

Patient demographics

Six demographic variables were extracted from each document, namely patient identifier, full name, date of birth, sex, postal code of residence, and document date. These fields served as identifiers for longitudinal record linkage and may be used as enriching input features (age, sex) for downstream prediction models.

Hospitalisation

For each inpatient episode, hospital admission and discharge dates were extracted. These dates defined encounter boundaries and established chronological ordering of clinical events. Hospitalisation dates served as temporal anchors when the exact date of a diagnosis was not documented.

Comorbidities

Pre-existing and incident comorbid conditions were extracted as diagnosis name and date of diagnosis pairs. Diagnoses were mapped to ICD-10 codes where available in the source text.

Primary diagnosis and staging

The primary oncological diagnosis was captured as a diagnosis name, date of first diagnosis, and histological subtype (e.g., IgG-kappa, IgA-lambda for myeloma). Disease staging was extracted separately for each staging assessment documented in the clinical record, comprising the staging system used (e.g., International Staging System, Revised ISS, Durie-Salmon), the assigned tumour stage, and the date of the staging assessment.

Cytogenetic assessment

Cytogenetic findings were extracted per reported abnormality, including the assessment date, the method used (fluorescence in situ hybridisation or conventional karyotyping), the affected gene (e.g., TP53, FGFR3), the genomic locus (e.g., 17p13, 4p16), and the type of detected abnormality (deletion, translocation, gain). Domain-specific high-risk markers including del(17p), t(4;14), t(14;16), gain(1q), and del(1p) were flagged for downstream risk stratification.

Bone marrow assessment

Each bone marrow evaluation was represented by the assessment date, the method used (aspiration, biopsy, or flow cytometry), the measured infiltration percentage, and the corresponding measurement unit.

Medication

Pharmacological treatments were extracted at the level of individual prescriptions, comprising the active substance name, dose, unit, administration schedule type (e.g., continuous, intermittent, as needed), and administration schedule pattern. The schedule pattern encoded either cycle-based regimens (e.g., days 1, 8, 15 of a 28-day cycle) for chemotherapy, or daily dosing distributions using the standard morning-noon-evening-night notation (e.g., 1-0-0 for once daily in the morning, 0-1-0-1 for noon and night administration) for supportive and chronic medications.

Laboratory measurements

Each laboratory result was extracted as a tuple of assessment date, analyte name, measured value, unit, and result status. The result status field indicated whether the measurement represented a standard validated result, point-of-care or preliminary measurement, or was flagged as invalid (e.g., haemolysed sample, insufficient material). Point-of-care and invalid results were retained in the extracted dataset but flagged for optional exclusion in downstream analyses.

Treatment lines

The course of disease was reconstructed as a sequence of treatment lines, each comprising treatment intent (first-line, salvage, or palliative), regimen name (e.g., VRd, KRd, Dara-VMP), treatment start and end dates, whether consolidation or maintenance therapy was administered, concurrent antiresorptive therapy (e.g., zoledronic acid, denosumab), dose reduction (yes/no), administration of high-dose therapy with autologous stem cell transplantation (yes/no), best response according to the International Myeloma Working Group uniform response criteria, progression-free survival (event occurrence, event date, and duration in days), and time to next treatment.

Refractory status

Drug-specific refractoriness was defined as non-response or disease progression on therapy, or within 60 days after discontinuation of a given agent [57]. Refractory status was captured for the following drug classes, including immunomodulatory agents (thalidomide, lenalidomide, pomalidomide), proteasome inhibitors (bortezomib, carfilzomib, ixazomib), monoclonal antibodies and targeted agents (anti-CD38 antibodies, elotuzumab, belantamab mafodotin, selinexor), and historical therapies (interferon, melphalan, cyclophosphamide, corticosteroids). Composite refractory categories were derived, including double-class (immunomodulatory agent plus proteasome inhibitor), triple-class (immunomodulatory agent plus proteasome inhibitor plus anti-CD38 antibody), quadruple-class, penta-drug (two immunomodulatory agents, two proteasome inhibitors, and an anti-CD38 antibody), and BCMA-directed therapy refractoriness.

Skeletal findings

Bone disease was captured through three entity types, namely osteolytic lesions (date of diagnosis and anatomical locations of individual lesions), fractures (date of finding, anatomical location, and any documented interventions with intervention dates), and general skeletal assessment findings. Lesion and fracture locations were recorded as free-text anatomical descriptors and normalised to a standardised skeletal region vocabulary during postprocessing.

Neurological complications

The presence or absence of polyneuropathy was extracted as a binary variable. Polyneuropathy is a frequent treatment-emergent adverse event associated with proteasome inhibitors and immunomodulatory agents and was tracked longitudinally across documents to monitor treatment tolerability.

SLiM-CRAB criteria

Myeloma-defining events according to the SLiM-CRAB framework were extracted as composite entities, specifically bone marrow infiltration exceeding 60% (S; assessment date, method, value), the presence of one or more osteolytic lesion on cross-sectional imaging (B; location, date), hypercalcaemia (C; date of documentation), renal insufficiency (R; date of documentation), and anaemia (A; date of documentation). The involved-to-uninvolved serum free light chain ratio (Li) was derived from extracted laboratory values where both measurements were available within the same assessment.

Death

Date of death was extracted where documented in the clinical record.

Information extraction pipeline

The information extraction pipeline comprised five stages operating sequentially on each clinical document (Extended Data Fig. 8).

In the first stage, complete discharge summaries, radiology reports, and pathology reports were converted from PDF into Markdown. Formatting cues such as font type, size, text positioning, and layout metadata were used to identify and exclude non-clinical content (hospital

contact information, page headers, footnotes, boilerplate disclaimers), while preserving document hierarchy including section headers, paragraph boundaries, and report metadata (Extended Data Fig. 8a).

The resulting normalised documents were then segmented into semantically coherent units, such as diagnosis, medical history, comorbidities, treatment descriptions, laboratory results, and radiology or pathology subsections, using rule-based heuristics that relied on heading keyword dictionaries, line-prefix patterns, and section delimiters. This hierarchical segmentation ensured that the downstream extraction model received only the most relevant textual context for each target entity, reducing noise and improving extraction quality (Extended Data Fig. 8d).

In the third stage, a locally deployed Llama 3.3-70B model performed template-driven extraction at the segment level, using a 4-bit AWQ-quantized variant (casperhansen/llama-3.3-70b-instruct-awq) to enable single-GPU deployment [58]. For each segment type, a dedicated template specified target fields, value constraints, and a strict JSON output schema (Listing 1). Invoking the model at the segment level rather than on full documents limited context size and focused extraction on relevant mentions. Templates were refined with domain experts to address ambiguous cases and improve consistency across reports. In addition to structured fields, the model optionally produced short narrative justifications describing the textual evidence supporting each extracted value.

Postprocessing and validation constituted the fourth stage. The model-generated narrative justifications were compared against the corresponding structured outputs to verify consistency between extracted values and their textual evidence. Discrepancies triggered automated rejection or re-querying of the model. Field-specific rule-based checks were applied for selected entities to exclude common confounders, for example differentiating osteolytic lesions from fractures in imaging-derived bone findings. This validation framework combined narrative consistency checks with targeted domain rules to improve robustness without requiring manual curation.

Finally, temporal extraction and clinical endpoint construction resolved all date expressions into DD-MM-YYYY format using regular-expression-based date discovery and rule-based resolution of relative temporal expressions anchored to the report date (Listing 2). When the day or month was not specified, missing components were imputed using January 1 of the corresponding year. Normalised events were sorted chronologically and grouped into lines of therapy using a second LLM pass that ingested the ordered event list and output structured treatment-line representations with normalised regimens. Progression-free survival and time to next treatment were computed from progression events and treatment-line boundaries, taking the minimum of the time to the next documented progression event or initiation of the subsequent line of therapy. Refractory status was determined using the rules defined in the entity definitions.

Listing 1. Example LLM extraction prompt and output schema for primary diagnosis segments

Prompt (abstracted): “Given the following segment from a clinical report, extract the requested fields. Output only valid JSON that conforms to the schema. If a field is not present, use null. For each extracted item include a short narrative quoting the minimal text that supports the extraction.”

Output schema (example):

```
{
  "segment_type": "primary diagnosis",
  "entities": [
    {
      "type": "Diagnosis",
      "narrative": "Free-text summary of the main oncologic diagnosis as stated in the letter.",
      "structuredDetails": {
        conditionName: "Multiple Myeloma" | other | -None- # if multiple myeloma is
        not mentioned specifically as the diagnosis, name the other
        histology: e.g. "IgA Kappa" | -None-
        dateOfFirstDiagnosis: "YYYY-MM" | "YY-MM" | "MM/YY" | "YYYY" | -None-
          # Date of initial Myeloma diagnosis.
          Typically date of ED.
          if "Aktuell:". use date of report
        tumorStagings:
        {
          system: str | -None- # e.g. R2-ISS, R-ISS, ISS, Durie-Salmon, etc.
          stage: str | -None- # e.g. "II"
          date: "YYYY-MM" | -None- # date of the staging,
          often dateOfFirstDiagnosis
        }
        [if applicable list other stagings mentioned]
      }
    }
  ]
}
```

Listing 2. Example regular expression for extraction of dates

```
(\b\d{4}\b) # year only, e.g., 2020
(\b\d{1,2}[/-]\d{1,2}[/-]\d{2,4}\b) # numeric dates: 15/07/2020 or 7-15-20
(\b(?:Jan|Feb|Mar|Apr|May|Jun|Jul|Aug|Sep|Oct|Nov|Dec)[a-z]*\.\.?s+\d{1,2},?\s+\d{4}\b)
# textual: July 15, 2020
```

Extraction performance evaluation

Pipeline accuracy was evaluated against manually annotated reference sets comprising 131 internal discharge summaries (18,875 structured outputs) and 14 external documents from four independent institutions (3,970 structured outputs). Accuracy was computed as the proportion of extracted attribute values matching reference annotations under exact-match criteria. For temporal attributes, two evaluation modes were applied, namely strict exact-match requiring agreement to the day and a lenient window accepting dates within one calendar month of the reference. For chronic conditions, the date of hospitalisation was accepted as a valid proxy for date of diagnosis when no more precise temporal anchor was available in the source document. When the pipeline could not recover a specific date, the hospitalisation date was assigned as a pragmatic temporal anchor to preserve longitudinal usability rather than omitting the attribute entirely. These cases are scored as errors under strict exact-match evaluation and reflect a deliberate design priority of structured output completeness over annotation fidelity. Restricting evaluation to predictions with complete date information increased mean date accuracy from 77.6% to 87.1%, confirming that residual temporal error originated primarily from ambiguous source documentation rather than extraction failure on well-specified dates.

All annotations were reviewed by three medically trained annotators comprising two medical doctoral candidates and one physician undergoing specialist training in radiology. Annotation followed a pre-specified schema. Disagreements were resolved by discussion, with the physician annotator serving as adjudicator for clinically ambiguous cases. Full per-entity accuracy results are reported in Supplementary Tables 4 and 5.

Diagnosis prediction: technical details

Sources of bias in longitudinal laboratory data

Four main sources of bias affect longitudinal laboratory data collected in routine clinical care. Testing frequency increases with disease severity and treatment complexity, resulting in overrepresentation of patients with advanced disease. Laboratory measurements are often ordered in response to clinical suspicion, introducing indication bias. Data originate from a tertiary university hospital, where case complexity is higher than in general care settings, and the cohort does not include healthy individuals. Predictions should therefore be interpreted within a clinical cancer population. Patients with longer survival contribute more observation time and thus more eligible laboratory sequences, introducing survival bias. The per-patient sequence cap described below partially mitigates this effect.

Cohort construction and filtering

Cohort construction, analyte selection, patient filtering, and dataset splits are described in Methods. The following paragraphs provide additional justification for the prediction horizon and per-patient sequence cap.

Choice of prediction horizon. A prediction horizon of 730 days was selected to capture the treatment-associated complication burden that accumulates over one to two lines of therapy. In relapsed multiple myeloma, median progression-free survival ranges from approximately 12 to 24 months depending on treatment line and regimen [52], and median time to next treatment follows a comparable distribution. In recurrent ovarian cancer, platinum-free intervals defining sensitivity thresholds and typical re-treatment cycles similarly fall within this range [8]. A 730-day horizon therefore aligns with the temporal scale over which clinically relevant complications are expected to manifest. Shorter horizons (e.g., 90 or 180 days) would preferentially capture acute toxicities such as neutropenic infections but miss slower-onset conditions. Longer horizons would introduce excessive noise from intercurrent treatment changes and disease progression events that alter the laboratory trajectory independently of the complications being predicted.

Maximum sequences per patient. The number of non-overlapping laboratory sequences sampled per patient was capped at ten for the general prediction task. Without this constraint, patients with prolonged treatment courses, who are also more likely to experience adverse outcomes, would contribute disproportionately to both training and evaluation, introducing survival bias and inflating performance estimates for endpoints correlated with treatment duration. A cap of ten sequences balances dataset size against this bias while preserving sufficient temporal coverage of each patient's disease course. For myelodysplastic syndromes, the sequence cap was increased to 20 per patient owing to low endpoint prevalence. At a cap of ten, positive sequences in the validation and test sets were insufficient for stable AUROC estimation. Increasing the cap to 20 yielded adequate positive case counts without materially altering the class balance of other diagnostic targets, as the additional sequences contributed primarily to patients with extended observation periods in whom therapy-related myeloid neoplasms are most clinically relevant. This design choice introduces additional survival bias for the MDS endpoint. Accordingly, MDS performance estimates should be interpreted with this enrichment in mind.

Model architectures and training configuration

Both the imputation and prediction models used a bidirectional transformer encoder based on the Qwen-3 architecture, trained from scratch without pretrained weights. Full architectural specifications and training hyperparameters are provided in Supplementary Tables 10 and 11.

Baseline models

Baseline methodology is described in Methods. For each laboratory sequence, analyte measurements were converted into bag-of-words count vectors over the analyte vocabulary. Logistic regression was trained with logistic loss. XGBoost was trained as a gradient-boosted decision tree classifier. Both baselines were evaluated using the same metrics (AUROC and enrichment ratio) and bootstrap procedure as the transformer.

Biomarker importance analysis

The masking approach is described in Methods. The masking effect for analyte a on diagnostic group g was defined as:

$$\Delta\text{AUROC}_{a,g} = \text{AUROC}_g^{\text{masked}(a)} - \text{AUROC}_g^{\text{full}}, \quad (1)$$

where $\text{AUROC}_g^{\text{full}}$ denotes the area under the receiver operating characteristic curve with all analytes present and $\text{AUROC}_g^{\text{masked}(a)}$ denotes the AUROC after masking analyte a . Negative values indicate that removing the analyte decreased predictive discrimination, with larger negative values reflecting greater importance.

Pairwise masking was performed analogously for all $\binom{40}{2} = 780$ analyte pairs. For each pair (a_1, a_2) and each diagnostic group, both analytes were simultaneously replaced with zero across all timesteps, and the pairwise masking effect was computed as:

$$\Delta\text{AUROC}_{(a_1, a_2), g} = \text{AUROC}_g^{\text{masked}(a_1, a_2)} - \text{AUROC}_g^{\text{full}}. \quad (2)$$

Table 10. Imputation model architecture and training configuration.

Component	Specification
Model backbone	Transformer encoder (Qwen-3, trained from scratch)
Hidden dimension	256
Number of layers	4
Number of attention heads	2
Number of Parameters	4229160
Token definition	One token per timestep encoding 39 analytes + age
Input projection	Linear projection of normalised analytes to hidden dim.
Prediction head	Linear layer mapping hidden dim. to analyte values
Loss function	Mean squared error on masked positions
Optimiser	Adam
Learning rate	0.001 (cosine annealing schedule)
Batch size	64
Input normalisation	Per-analyte z-score normalisation (training set stats)
Missing value handling	Zero imputation with analyte-specific presence mask
Early stopping	MSE plateau (≥ 40 epochs)
Typical training duration	~ 600 epochs
P (mask individual analyte)	0.2
P (mask entire timestep)	0.05

Table 11. Diagnosis prediction model architecture and training configuration.

Component	Specification
Model backbone	Transformer encoder (Qwen-3, trained from scratch)
Hidden dimension	64
Number of layers	4
Number of attention heads	2
Number of Parameters	484707
Token definition	One token per timestep encoding 39 analytes + age + sex (single token per sequence)
Input projection	Linear projection of normalised analytes to hidden dimension
Token aggregation	Attention pooling across all token-wise last-layer transformer representations.
Prediction head	Linear layer applied to aggregated last-layer tokens
Loss function	focal loss, $\gamma 2.0$ (class weights from training set)
Optimiser	Adam
Learning rate	0.001 (cosine annealing schedule)
Batch size	64
Input normalisation	Per-analyte z-score normalisation (training set statistics)
Missing value handling	Pre-imputation via imputation model
Early stopping	AP plateau (≥ 20 epochs)
Typical training duration	~ 40 epochs
Prediction horizon	730 days
Cross-validation	5 patient-level folds
Ensemble strategy	Mean logits across fold-specific models

Evaluation metrics

Model performance was evaluated using diagnosis-level classification metrics computed independently for each target diagnosis and subsequently aggregated across diagnostic groups.

Area under the receiver operating characteristic curve. Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC), which quantifies the probability that a randomly selected positive instance receives a higher predicted score than a randomly selected negative instance. AUROC is threshold-independent and invariant to prevalence, which makes it comparable across endpoints but also insensitive to the precision achievable at any given prevalence. It is widely reported in the clinical prediction literature, facilitating comparison with existing models.

Average precision. To complement AUROC, we additionally computed average precision (AP) as implemented in `scikit-learn`. Whereas AUROC summarizes ranking performance over both classes and is invariant to prevalence, AP summarizes the precision–recall curve and therefore characterizes performance among the highest-ranked, positively-predicted patients, the operating region most relevant when positive cases are rare. AP is sensitive to prevalence: a model may attain a high AUROC while achieving low precision among its top-ranked predictions, a discrepancy that AUROC cannot reveal but AP makes explicit. Reporting both metrics thus separates a model’s ability to rank cases (AUROC) from its ability to concentrate true cases among its most confident predictions (AP). AP is computed as a weighted sum of precision values at observed operating points, with weights given by the increase in recall between successive thresholds:

$$AP = \sum_{k=1}^K (R_k - R_{k-1}) P_k, \quad (3)$$

where P_k and R_k denote the precision and recall computed at the k -th distinct decision threshold after sorting predictions in descending order, and $R_0 = 0$. This formulation computes a non-interpolated approximation of the area under the precision–recall curve using a step-function summation, which avoids the overestimation associated with linear (trapezoidal) interpolation. Average precision provides a threshold-independent performance summary and is appropriate in settings with substantial class imbalance.

The enrichment ratio, defined as AP divided by endpoint prevalence, quantifies the gain in top-ranked precision relative to a random classifier, whose expected AP equals the prevalence. Values above one therefore indicate that true cases are concentrated among the highest-ranked patients more than chance would predict, which is informative under severe class imbalance, where a moderate AUROC can still correspond to substantial enrichment.

Diagnostic group-level evaluation. For predefined diagnostic groups comprising multiple related diagnoses, group-level prediction scores were constructed by taking the maximum predicted probability across all constituent diagnoses for each sample:

$$\hat{y}_g = \max_{d \in g} \hat{y}_d. \quad (4)$$

Group-level ground-truth labels were defined analogously:

$$y_g = \max_{d \in g} y_d. \quad (5)$$

AUROC and average precision were then computed for each diagnostic group using these derived predictions and labels.

Confidence intervals. Uncertainty in performance estimates was quantified using nonparametric bootstrap resampling of the held-out test set. The test set was resampled with replacement at the sequence level 1000 times, and AUROC and average precision were recomputed for each diagnosis and diagnostic group on each replicate. Performance was summarised as the median across replicates, and confidence intervals were obtained from the empirical bootstrap distribution using percentile-based intervals.

Unless stated otherwise, all metrics were computed on the held-out test set at the diagnosis level and reported either individually or as macro-averaged values across diagnoses.

Dataset summary

Dataset splits and cohort sizes are reported in Methods. Per-endpoint prevalences for all 162 individual ICD codes and diagnostic groups are provided in Supplementary Tables 6 and 7.