Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Generative Multi-State Deep Learning Framework for Dynamic Virtual-Twin Modeling of Multiple Myeloma

Sebastian Schwick
Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Yasamin Salimi
Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Shammi More
Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Guillaume Azarias
TriNetX Oncology GmbH
Holger Fröhlich
holger.froehlich@scai.fraunhofer.de

Fraunhofer Institute for Algorithms and Scientific Computing SCAI

Article

Keywords: Multi-state models, generative AI, multiple myeloma, virtual twins

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11204103/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Generative Multi-State Deep Learning Framework for Dynamic Virtual-Twin Modeling of Multiple Myeloma

Sebastian Schwick [1], Yasamin Salimi[1] , Shammi More[1], Guillaume Azarias[2], and Holger Fröhlich [1,3,4,5] on behalf of the CERTAINTY consortium

1. Department of Biomedical AI & Data Science, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany; 2. TriNetX Oncology GmbH, Freiburg, Germany; 3. Bonn-Aachen International Center for Information Technology (b-IT), University of Bonn, Bonn, Germany; 4. Institute for Digital Medicine, University Hospital Bonn, Bonn, Germany; 5. Lamarr Institute for Machine Learning and Artificial Intelligence, Dortmund, Germany

Keywords: Multi-state models, generative AI, multiple myeloma, virtual twins

Corresponding author: Holger Fröhlich (holger.froehlich@scai.fraunhofer.de)

Abstract: Multiple myeloma (MM) is a heterogeneous and incurable cancer of plasma cells in which patients experience repeated disease relapses and widely variable survival outcomes. Existing risk prediction models focus primarily on patients at initial diagnosis, leaving a critical gap for individuals with relapsed or refractory disease, where previous treatments could potentially influence future outcomes. Computational virtual twins (VTs) represent a promising approach for personalized medicine by continuously updating patient-specific predictions as new clinical information becomes available. However, enabling such personalized prediction requires artificial intelligence (AI) models that can jointly represent disease trajectories and heterogeneous, longitudinal patient data.

We present DG-MSM, a generative multi-state deep learning framework designed as the computational foundation for VT modeling. The model jointly learns disease-state transitions, event timing, and heterogeneous longitudinal patient characteristics, incorporating the full patient history into the risk assessment. Beyond dynamic risk prediction, this generative structure enables uncertainty-aware, patient-specific simulation of future disease trajectories, allowing outcomes to be examined as full predictive distributions rather than single risk scores. DG-MSM achieves competitive performance relative to dedicated time-to-event models, while providing long-term survival predictions that grow more accurate as a patient's treatment history accumulates, an advantage most pronounced in later treatment lines. These findings demonstrate the potential of DG-MSM to enable predictive virtual twins in MM and other chronic diseases, providing dynamically updated prognosis, individualized risk assessment, and patient-specific trajectory simulation to support treatment planning.

Introduction

Multiple myeloma (MM) is a heterogeneous and incurable cancer characterized by complex disease trajectories, repeated relapses, and substantial variability in patient outcomes [1]. While accurately predicting these outcomes is essential for treatment planning [2], classical risk stratification approaches for MM assign patients to broad risk groups at diagnosis without predicting individual outcomes or accounting for the impact of specific treatment sequences [3–5]. Deep learning methods have demonstrated the ability to overcome these limitations by incorporating high-dimensional, multimodal data, including clinical variables, genomics, and longitudinal biomarkers [6]. The recently proposed Individualized Risk Model for MM (IRMMa) represents a significant advance toward individualized prognosis in newly diagnosed MM (NDMM) by integrating clinical, genomic, and treatment data within a multi-state framework [7]. However, IRMMa's predictions are conditioned on covariates assessed at diagnosis and do not update as the disease evolves. More broadly, current risk assessment remains largely focused on NDMM, despite management of relapsed or refractory MM (RRMM) being fundamentally shaped by the patient's treatment history, making individualized, dynamically updated prediction in later lines of therapy (LoT) a critical open challenge [8,9].

Computational virtual twins (VTs) are a promising conceptual framework for addressing this challenge [10,11]. Within the CERTAINTY consortium, a VT is defined as an in silico system integrating multiple patient-specific computational models across biological scales to simulate disease and treatment trajectories [12]. Weirauch et al. identify data-driven artificial intelligence (AI) models as essential components of VT architectures, with applications such as clinical decision-support software that guides optimal treatment sequences based on patient-specific probabilistic assessments. At the disease-trajectory level, this requires models that can represent sequential state transitions, generate patient-specific predictions conditioned on evolving longitudinal data, and quantify predictive uncertainty. MM patients progress through sequential LoTs [13], face competing risks of relapse and death at each stage, and accumulate longitudinal clinical histories that increasingly inform the prediction of possible outcomes. Consequently, a dynamic

VT model would be particularly valuable for supporting the repeated treatment decisions that define long-term MM management.

A variety of statistical and AI-based approaches have been developed to address distinct aspects of this challenge. Deep learning survival models such as DeepHit [14] and CoxTime [15] have advanced discriminative performance for time-to-event prediction. Variational autoencoder (VAE)-based approaches like CR-SAVAE [16] instead jointly learn a generative model of covariates and time-to-event outcomes, deriving parametric, per-risk survival distributions from a shared latent representation. However, these methods are designed for the prediction of individual events or competing risks rather than modeling complete multi-state trajectories. Classical multi-state models (MSMs) [17,18] provide a theoretical framework for sequential transitions, yet typically require Markov or semi-Markov assumptions and proportional hazards for each transition, limiting their flexibility for high-dimensional, heterogeneous, and irregularly sampled longitudinal data. Generative approaches for clinical data, including heterogeneous incomplete variational autoencoder (HI-VAE) [19] and MultiNODEs [20], have demonstrated the ability to produce realistic synthetic patient data with heterogeneous feature types, although they were not designed for event-time modeling or multi-state dynamics. Recent large language model (LLM)-based approaches have been proposed for laboratory value forecasting [21,22] in the context of VTs, but they address different modeling tasks than generative trajectory simulation with calibrated uncertainty. While each of these directions advances an important aspect of patient modeling, developing flexible AI-driven strategies that jointly predict patient trajectories across multiple disease states while learning the underlying data-generating process remains a significant methodological gap.

To address this gap, we propose Dynamic Generative Multi-State Model (DG-MSM), a deep generative multi-state framework that jointly learns the full disease trajectory, including state transitions, event times, and heterogeneous longitudinal covariates. Compared to existing approaches that either fit models to individual events or require strong parametric assumptions about transition hazards, DG-MSM combines variational inference with a discrete-time multi-state jump process. This enables history-aware

predictions while simultaneously learning a longitudinal latent patient representation. As a fully generative model, DG-MSM enables not only individualized dynamic risk prediction that improves as patient history accumulates, but also patient-specific trajectory simulation and synthetic data generation. We evaluate DG-MSM on two distinct datasets - a real-world MM database and an observational clinical study. We demonstrate competitive discriminative performance relative to dedicated survival models and patient-specific calibrated prediction intervals that can be updated as new data become available. These results establish DG-MSM as a computational foundation for disease-trajectory modeling within VT architectures.

Results

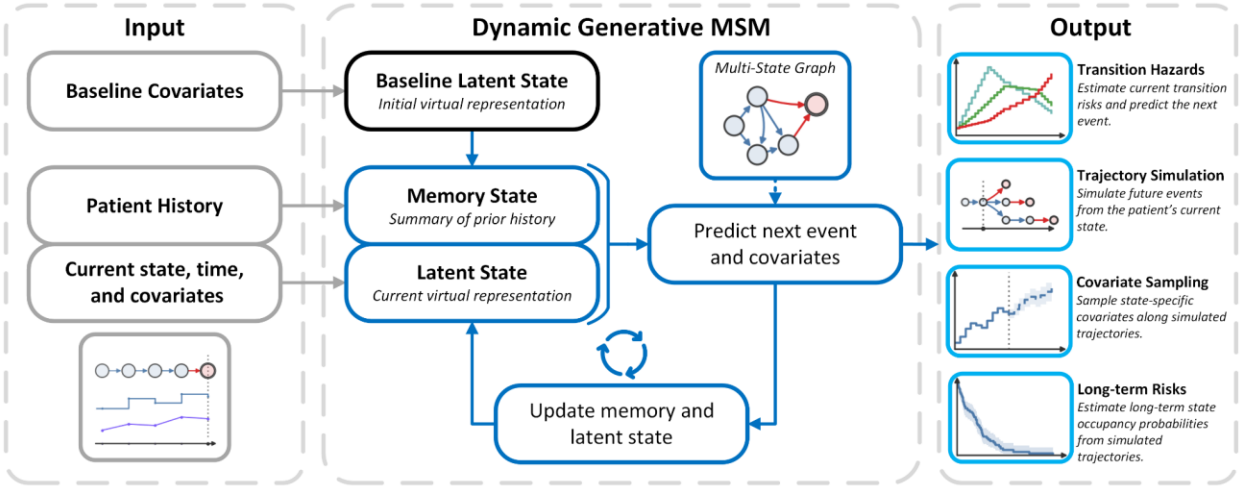


Fig. 1: Schematic overview of DG-MSM as a computational foundation of a VT.

DG-MSM as a Flexible Model Architecture for Multi-State Disease Processes

We propose DG-MSM, a generative deep learning model for simulating complete patient disease trajectories within a clinically defined multi-state process **(Fig. 1)**. DG-MSM jointly models the sequence and timing of clinical events and the covariates observed in each state, conditioned on the patient's entire medical history up to that point. The model consists of three components: a HI-VAE for baseline covariates, a recurrent variational autoencoder for longitudinal visit data, and a prediction head for Multi-State Risks in Discrete Time (MuStaRDT). The baseline HI-VAE module encodes static patient

characteristics into a latent representation. This representation initializes the recurrent encoder, which sequentially processes the clinical observations accumulated along the trajectory, building a latent memory that integrates information from the full observed history without requiring Markov assumptions. The latent memory representation, together with the current clinical observations and the patient's baseline latent characteristics, is passed to the multi-state transition head, which outputs probability distributions over reachable event types and their time-to-occurrence. The model handles heterogeneous data types (e.g., continuous, categorical, and ordinal) and missing values via feature-specific likelihood functions. For further details on the architecture and training procedure, we refer to the Methods section. Evaluation of synthetic data generated by conditioning on baseline covariates is shown in the supplementary material (Supplementary Fig. 1 and 2).

MSM Design Based on Data Characteristics

We validated the model's risk predictions using two individual datasets representing different data types collected in this field and compared its performance against three previously established models (**Table 1**). While the CoMMpass study (NCT01454297) is an observational clinical study with detailed response assessments, TriNetX provides treatment-focused real-world data. Following the preprocessing steps and covariate selection described in the Methods section, a total of 1,008 patients from CoMMpass and 8,878 patients from TriNetX were included in the analysis.

In MM, treatment is structured in sequential LoTs, as the disease is currently considered incurable and patients may require subsequent treatments due to disease persistence, relapse, progression, or inadequate response to prior therapies. Each subsequent line therefore reflects a new clinical context shaped by prior treatment exposure, treatment response, disease status, and available therapeutic options. Accordingly, we defined the MSM using LoTs as clinically meaningful treatment phases representing different stages of the patient journey. This design enables dynamic estimation of survival risks based on the patient's current treatment context and aims to support individualized risk assessment at clinically relevant decision points.

We defined dataset-specific multi-state graphs that appropriately reflect the characteristics of each setting (**Fig. 2**). This approach accounts for differences in data availability, follow-up duration, and clinical granularity between the two datasets. For TriNetX, five LoT states were included because later treatment lines contained substantially fewer patients and transition events (Supplementary Fig. 3), limiting reliable estimation of transition probabilities. Thus, restricting the graph to five LoT states allowed robust estimation while preserving clinically relevant treatment trajectories. For CoMMpass, the multi-state structure reflects the available clinical follow-up and response assessments, such that the sojourn time (i.e., the duration a patient spends in a state before transitioning) in the initial state corresponds to progression-free survival (PFS), and the sojourn time in the second state represents survival following the first disease progression (Supplementary Fig. 4).

Summary statistic	**CoMMpass**	**TriNetX**
Number of patients	1,008	8,878
Age, median (mean)	63 (62.85)	71 (70.01)
Female, n (%)	400 (39.7%)	3592 (40.5%)
ISS I at baseline, n (%)	375 (37.2%)	1524 (17.2%)
ISS II at baseline, n (%)	365 (36.2%)	4025 (45.3%)
ISS III at baseline, n (%)	268 (26.6%)	3329 (37.5%)
Median follow-up, years	6.65	3.16
Median OS, years	8.53	7.07
Deaths, n (%)	296 (29.4%)	1856 (20.9%)

Table 1: Summary statistics of the CoMMpas and TriNetX datasets. ISS: International Staging System. OS: Overall survival.

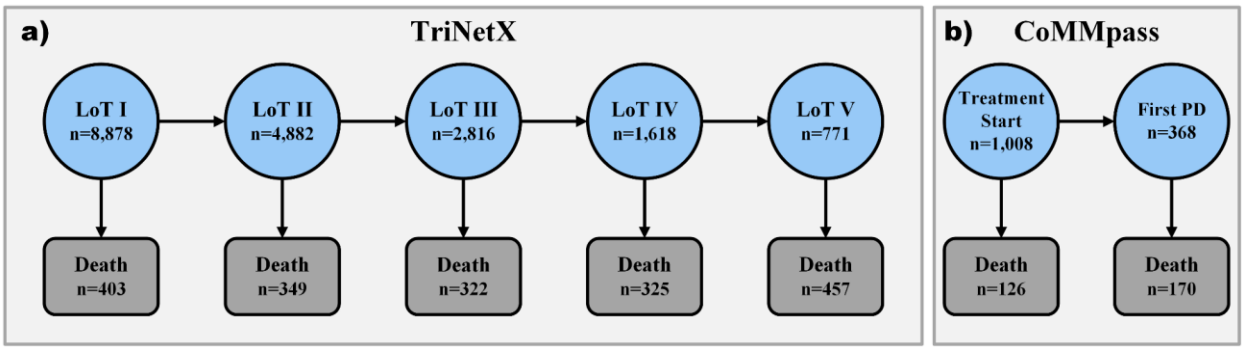


Fig. 2: The multi-state graphs for the two use cases: **a)** the TriNetX dataset and **b)** the CoMMpass dataset. (*PD*: progressive disease).

State-Specific Risk Assessment

We evaluate the risk prediction performance of the proposed model by estimating the individual-level sojourn time in the current state. Specifically, for each entry into a non-terminal state, we treat the subsequent time until the next transition or censoring as a right-censored outcome and evaluate each state-visit as an individual time-to-event setting. This allows a direct comparison between MSMs and standard time-to-event models that predict classical survival distributions. We compare DG-MSM against three baseline models representing distinct modeling paradigms: neural non-proportional hazard models, discrete-time competing-risks models, and classical multi-state regression. CoxTime [15] represents neural Cox-based models with non-proportional, non-linear hazards, while DeepHit [14] provides a discrete-time competing-risks model closest in structure to DG-MSM's per-state transition head. For both, we fit separate state-specific models predicting sojourn time in each non-terminal state, with DeepHit treating possible outgoing transitions as competing events. As the classical statistical baseline, we consider a semi-Markov Cox-based MSM (i.e., using time since entry into the current state as the time scale) with transition-specific baseline hazards and ridge penalization, implemented using the mstate R package [17]. Unlike the per-state CoxTime and DeepHit models, DG-MSM and the Cox-MSM each fit a single joint model over complete multi-state trajectories. We assess the state-specific risk predictions using the concordance index (C-index) [23], with and without inverse probability of censoring weighting (IPCW),

time-dependent area under the receiver operating characteristic curve ($AUC(t)$), and integrated Brier score (IBS).

Use Case 1: Real-World Data (TriNetX)

Per-state risk prediction on the TriNetX dataset was assessed using 5-fold cross-validation (CV) stratified by each patient's last observed LoT. Among the evaluated metrics, IBS is particularly informative for sojourn-time prediction because it summarizes prediction error over time. DG-MSM achieves the best mean performance across states for all metrics (C-index 0.697, IPCW C-index 0.682, IBS 0.102, and mean $AUC(t)$ 0.739; Supplementary Table 1). It also yields the lowest IBS in every LoT, suggesting the most accurate sojourn-time prediction (**Fig. 3d**). The only exception to DG-MSM's otherwise leading performance occurred at LoT 1, where CoxTime achieved marginally higher concordance (C-index 0.682 vs. 0.677, **Fig. 3a**; IPCW C-index 0.664 vs. 0.658, **Fig. 3b**), although DG-MSM already achieved the lowest IBS and highest $AUC(t)$ at this line. From LoT 2 onward, DG-MSM led across all four metrics.

Overall, the deep learning models outperform the classical Cox-MSM in most states, which may reflect greater flexibility to capture non-linear covariate effects and non-proportional hazards. CoxTime and DeepHit followed similar trends across treatment lines, with CoxTime consistently achieving somewhat higher discrimination and lower prediction error. DG-MSM's advantage becomes more apparent in later treatment lines, most notably in LoT 5, where it achieves an IPCW-adjusted C-index of 0.758 and an $AUC(t)$ of 0.852, compared with 0.698 and 0.773 for CoxTime (**Fig. 3c**). This pattern is consistent with the DG-MSM architecture, in which recurrent history encoding may accumulate clinically relevant information across treatment lines.

Across most models, IBS increases in later LoTs, consistent with the smaller cohort size and higher censoring at these states (Supplementary Fig. 3). DG-MSM and Cox-MSM instead showed a decrease in IBS at LoT 5, which may partly reflect that LoT 5 permits only a single outgoing transition to death rather than the competing progression/death choice present at earlier lines. Cox-MSM showed a similar gain in concordance at LoT 5 (C-index from 0.596 to 0.696, the largest single-line change of any model), while CoxTime

and DeepHit improved only modestly (+0.045 and +0.042). This shared behavior between the two joint MSMs suggests the effect comes from LoT 5's simplified transition structure rather than from DG-MSM's recurrent architecture specifically.

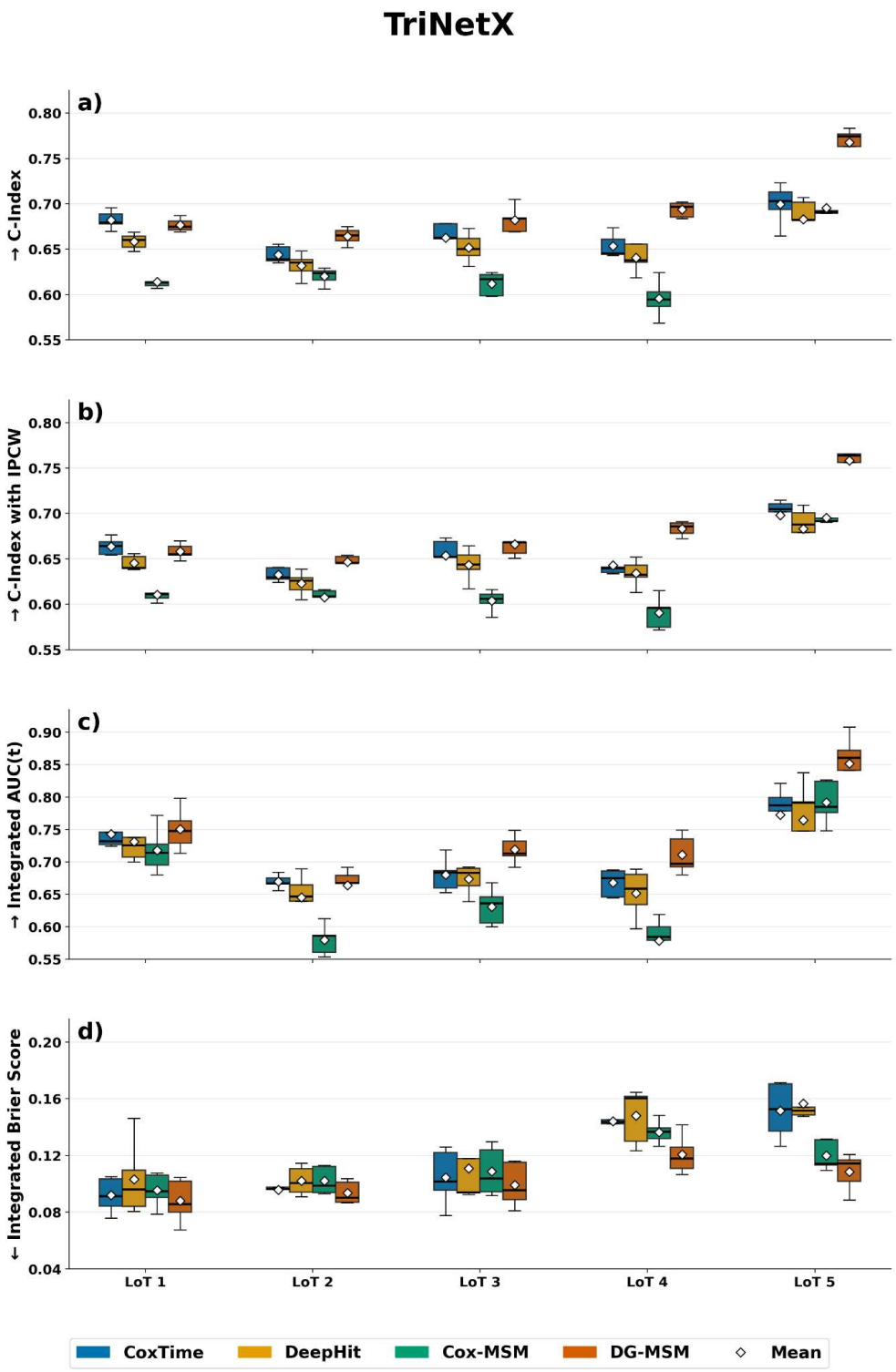


Fig. 3: Results for 5-fold nested CV using the TriNetX dataset. **a)** C-index, **b)** C-index with IPCW, **c)** integrated AUC over time, **d)** integrated Brier score (IBS). Please note that lower IBS indicates better calibration and discrimination, whereas for **a)** - **c)** a higher value indicates stronger discrimination.

Use Case 2: Study Data (CoMMpass)

As an additional validation of our framework, we repeated the 5-fold CV experiment on the CoMMpass clinical dataset. This use case assesses the applicability of the proposed models to a dataset with a different structure and clinical data representation than those used in the previous analysis. Model performance was evaluated using the same metrics as in the TriNetX use case. At treatment start, where no trajectory history exists yet, DG-MSM, Cox-MSM, and CoxTime performed comparably across all four metrics (**Fig. 4,** e.g., C-index: 0.678, 0.680, and 0.670, respectively), while DeepHit performed worse throughout. DG-MSM achieved a better $AUC(t)$ (0.723 vs. 0.721 for Cox-MSM, **Fig. 4c**) but a slightly higher IBS than Cox-MSM (0.186 vs. 0.178, **Fig. 4d**). The comparable performance across models at this state is consistent with DG-MSM's recurrent encoder having no accumulated history to utilize.

After the first disease progression, DG-MSM's discriminative advantage became clearer. It achieved the highest C-index among all models, both without (0.741 vs. 0.725 for CoxTime and 0.718 for Cox-MSM, **Fig. 4a**) and with IPCW (0.696 vs. 0.687 and 0.665, respectively, **Fig. 4b**). CoxTime, however, became a considerably stronger baseline at this state than at treatment initiation, matching Cox-MSM's $AUC(t)$ (0.734, both models) and achieving a marginally lower IBS than DG-MSM (0.205 vs. 0.208). DG-MSM nonetheless retained the highest $AUC(t)$ among the multi-state approaches at this state (0.743) and continued to outperform DeepHit on every metric in both states. Estimates after first disease progression also carried substantially greater cross-fold variability than at treatment initiation (e.g., C-index standard deviation 0.036-0.067 vs. 0.027-0.035, Supplementary Table 2), consistent with the smaller number of patients and observed events available for this state (n=368, 53.8% censored, Supplementary Fig. 4).

Taken together, DG-MSM and Cox-MSM consistently outperformed DeepHit, showing that explicit modeling of disease progression structure can provide measurable gains even in this simpler, two-state setting. DG-MSM's own advantage over Cox-MSM was concentrated specifically in discrimination after the first recorded transition, while its prediction error (i.e., IBS) was modestly higher than Cox-MSM's, and after first disease

progression also higher than CoxTime's, indicating that its discriminative gains in this smaller, structurally simpler setting are not fully mirrored in overall prediction error. DG-MSM nonetheless exhibited the lowest cross-fold IBS variance among all deep learning models, indicating stable predictions across patient subgroups. These results support the hypothesis that longitudinal history encoding becomes increasingly valuable as patients accumulate clinical events and show that a single generative model remains competitive with, and in several respects exceeds, dedicated survival models even in a comparatively low-data, low-complexity setting.

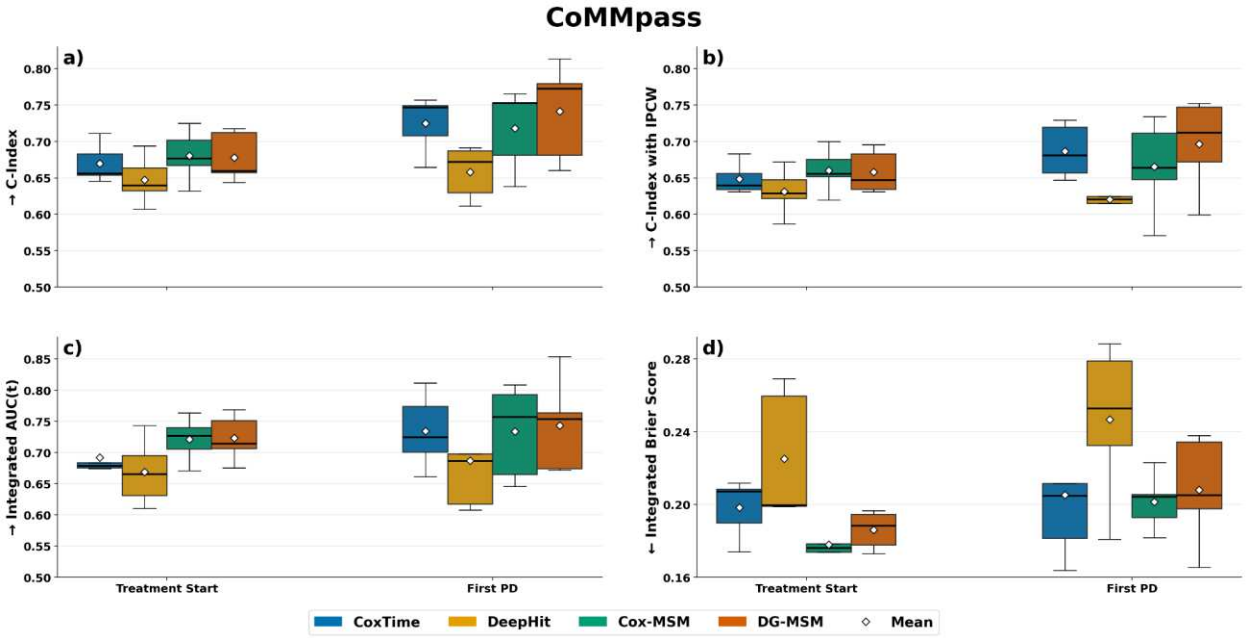


Fig. 4: Results for 5-fold nested CV using the CoMMpass dataset. **a)** C-index, **b)** C-index with IPCW, **c)** integrated AUC over time, **d)** integrated Brier score. Please note that lower IBS indicates better calibration and discrimination, whereas for **a)** - **c)** a higher value indicates stronger discrimination.

Patient-specific Simulations

Beyond discriminative risk prediction, DG-MSM's generative structure simulates a distribution of future disease trajectories for each patient, conditioned on their clinical history. This is implemented by repeatedly sampling possible future trajectories from the model, conditioned on each patient's history up to their current state, from which any outcome of interest, such as overall survival (OS), can be estimated as described in

Methods. We evaluated whether the resulting predictive distributions are themselves statistically trustworthy, and how the resulting point estimates of residual OS compare to a conventional discriminative benchmark as more history accumulates. We assessed this directly on the TriNetX dataset, using the same five-fold CV splits as above. We focused on treatment initiation (LoT 1), which comprises the full test cohort in every fold and represents the most demanding case in the sequence, since OS predictions from this state require the longest, most compounded simulated horizon of any landmark (full calibration results are shown in Supplementary Fig. 5-8). For each patient, we sampled 1,000 future trajectories conditional on baseline and current-state covariates. We assessed distributional (D-)calibration [24] and the prediction interval coverage probability $PICP(\alpha)$ [25], pooled across all five folds. The $PICP(\alpha)$ was adapted to right-censored outcomes via IPCW restricted to patients with an observed event.

D-calibration for OS showed a deviation from uniformity that was small in absolute magnitude, with a maximum cumulative distribution function (CDF) deviation of 0.030 **(Fig. 5h)**, indicating that observed deaths occurred slightly later than implied by each patient's predicted distribution. Given the large pooled sample size (N = 8,878), even this modest deviation was formally detectable ($\chi^2 = 48.0$, $p = 2.5 \times 10^{-7}$). Expressed as the D-calibration effect size $R(\theta)$[26], this deviation was small ($R(\theta) = 5.4 \times 10^{-4}$, corresponding to Cohen's $w = 0.07$[27]). Prediction-interval coverage showed the same pattern: the $PICP(\alpha)$ exceeded nominal levels throughout, consistent with intervals that are wider than strictly required rather than falsely narrow **(Fig. 5g)**.

For comparison, sojourn time showed no detectable deviation from uniformity ($\chi^2 = 5.5$, $p = 0.79$) and a substantially smaller effect size than OS ($R(\theta) = 6.2 \times 10^{-5}$, Cohen's $w = 0.02$), with maximum CDF deviation 0.005 (**Fig. 5d**). The $PICP(\alpha)$ for sojourn time in LoT 1 also exceeded nominal levels throughout, by a smaller margin than for OS **(Fig. 5c)**. The modest deviation in OS likely reflects error accumulation across recursive transitions, as OS prediction propagates uncertainty across multiple transitions, unlike sojourn time, which involves a single event. Together, these results show that DG-MSM's simulated

trajectories are calibrated, patient-specific predictive distributions, and that this calibration is preserved across the recursive simulation required to generate long-term trajectories.

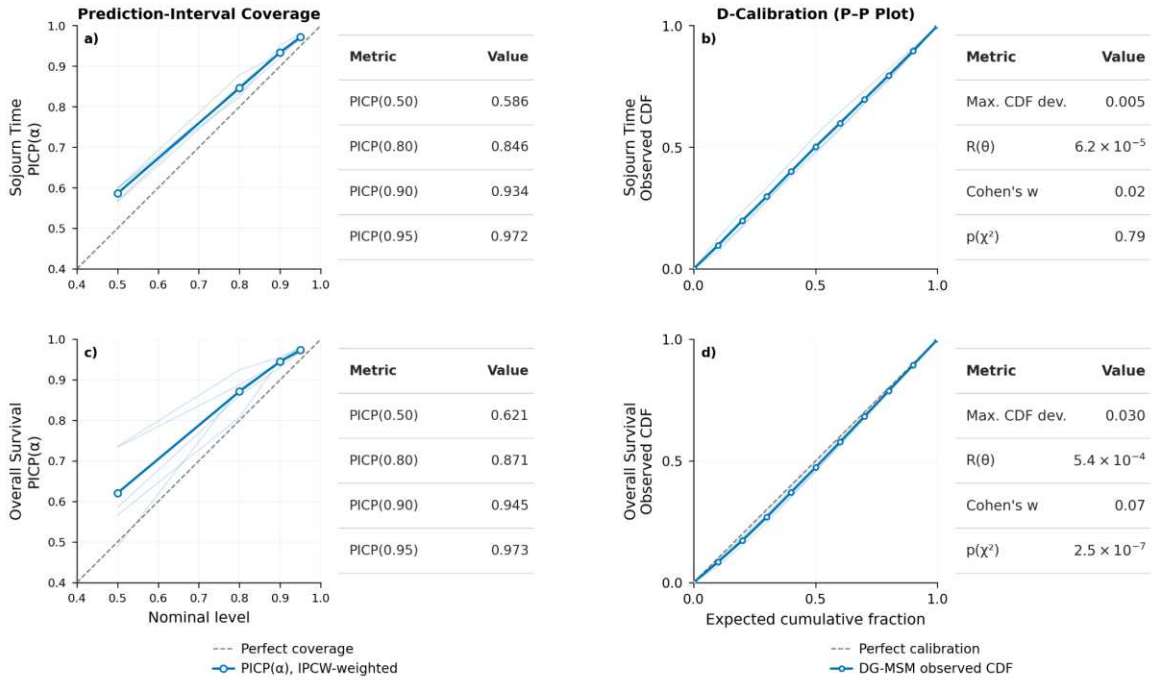


Fig. 5: Calibration of DG-MSM's simulation-based predictions (5-fold cross-validation). Probabilistic calibration of simulated trajectories at treatment initiation (LoT 1), pooled across all five folds: $PICP(\alpha)$ for **a)** sojourn time and **c)** OS; D-calibration plots for **b)** sojourn time and **d)** OS**.**

Dynamic Updating of Risk Predictions

Building on the simulation-based residual OS estimates introduced above, we next evaluated how discrimination for residual OS changes as additional treatment history accumulates along the patient's trajectory. In this analysis, we focused on the TriNetX dataset, which provides longer follow-up across later treatment lines, and results were generated over the same 5-fold CV splits used in the previous analyses. For each patient in the test split and observed LoT, we used this simulation procedure to estimate residual OS at each state, as described in Methods. As a comparative benchmark, we trained a separate CoxPH model with elastic-net regularization for each LoT using the same outer CV splits and hyperparameter optimization procedure. CoxPH therefore provides a

conventional comparator that estimates residual survival from the information available at the current LoT **(Fig. 6)**.

In earlier treatment lines, where longitudinal information remains limited, CoxPH was competitive and, in some cases, performed slightly better than DG-MSM. At LoT 1, CoxPH achieved a higher integrated $AUC(t)$ score than DG-MSM (0.682 vs. 0.667) and a lower IBS (0.157 vs. 0.169), with a similar pattern observed at LoT 2. This pattern is consistent with the simpler CoxPH model yielding more stable long-range risk rankings, whereas DG-MSM propagates uncertainty through repeated stochastic simulation of future trajectories. From LoT 3 onward, DG-MSM consistently outperformed CoxPH on discrimination, and the gap widened at later treatment lines. Integrated $AUC(t)$ increased from 0.737 at LoT 3 to 0.769 at LoT 4 and 0.852 at LoT 5 for DG-MSM, compared with 0.702, 0.688, and 0.793 for CoxPH. The same trend was observed for concordance, with DG-MSM achieving IPCW C-index values of 0.702 at LoT 3, 0.720 at LoT 4, and 0.758 at LoT 5, compared to 0.653, 0.641, and 0.696 for CoxPH. Prediction error showed a similar pattern: CoxPH had marginally lower IBS through LoT 3 (e.g., 0.164 vs. 0.168 at LoT 3), but DG-MSM performed better from LoT 4 onward, with the largest improvement at LoT 5 (0.108 vs. 0.163). At LoT 4 and LoT 5, DG-MSM outperformed CoxPH across all four metrics.

Overall, these findings suggest that modeling the evolving treatment trajectory is most beneficial once sufficient longitudinal history is available. The widening gap between DG-MSM and CoxPH at later LoTs indicates that the sequence of prior treatment states contains additional prognostic signals that improve residual OS prediction compared with models based only on the current LoT.

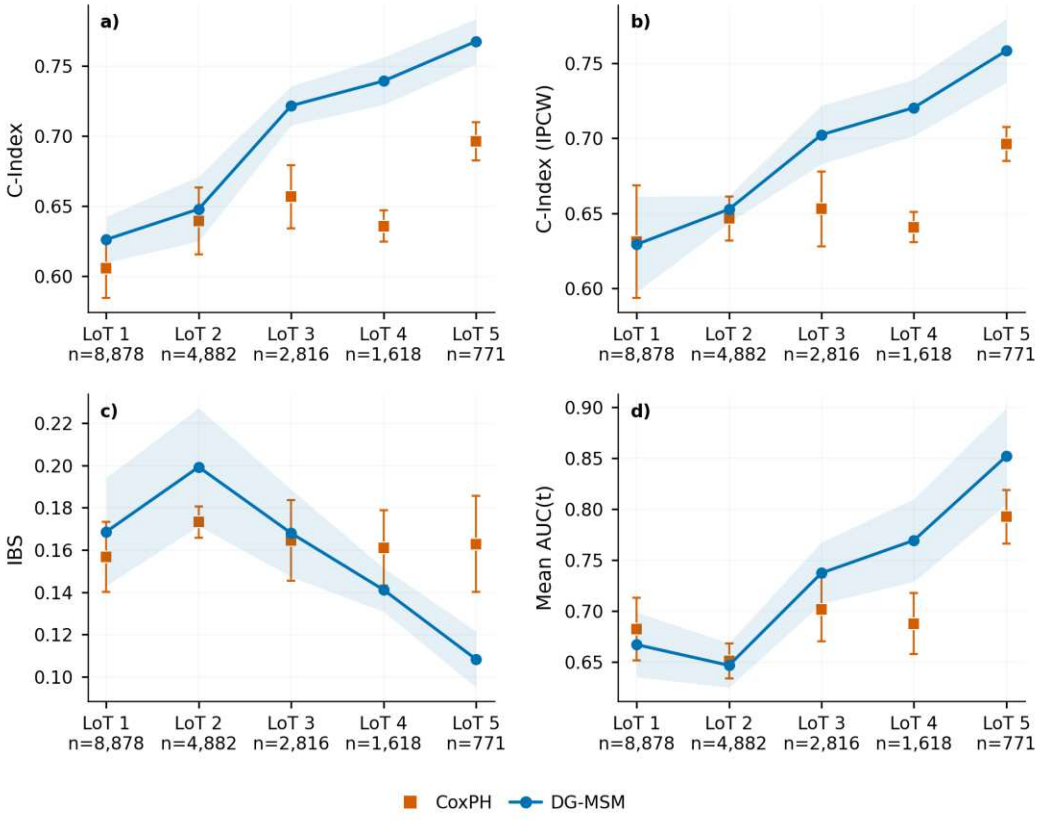


Fig. 6: Performance of DG-MSM's simulation-based predictions. Discrimination and prediction-error metrics for OS at each evaluated LoT with mean ± standard deviation: **a)** C-index, **b)** C-index with IPCW, **c)** integrated AUC(t), and **d)** IBS.

Discussion

In this work, we presented DG-MSM, a dynamic generative MSM that jointly addresses multi-state transitions, generative modeling of heterogeneous incomplete covariates, and long-term survival simulation within a single framework. The model supports state-specific risk assessment, dynamically updated risk predictions, and patient-specific forward simulation with empirically validated calibration of its simulated uncertainty. From a clinical perspective, DG-MSM complements established staging systems (ISS [3], R-ISS [4]) by incorporating longitudinal trajectory information that static baseline assessments, by design, do not capture.

The discrete-time categorical formulation avoids restrictive distributional assumptions on transition hazards, while the recurrent memory state relaxes the Markov assumption by

allowing transitions to depend on the full observed history rather than only the current state. Unlike modular approaches that chain independent imputation, prediction, and generation steps, DG-MSM optimizes all components jointly within a single variational inference framework. This ensures that the learned patient representation is designed to capture those features that are simultaneously informative for state transitions, covariate reconstruction, and trajectory simulation. As a result, DG-MSM outperformed dedicated per-state models (e.g., CoxTime and DeepHit) in the majority of evaluated cases and achieved comparable performance in the remaining cases, while capturing non-linear, non-proportional covariate effects that the classical Cox-MSM cannot represent.

On the TriNetX dataset, DG-MSM achieved the best mean performance across all states, with its discriminative advantage growing progressively from early treatment lines to later lines, where the accumulated trajectory history provided increasingly informative conditioning signals. This pattern is consistent with the architectural hypothesis that a recurrent encoder that accumulates patient history becomes most valuable where treatment decisions are increasingly difficult and where simpler models lack access to the sequential history needed to leverage prior disease dynamics.

On the CoMMpass dataset, DG-MSM demonstrated robust generalization to a lower-resource setting with shorter trajectory depth (two non-terminal states), achieving the highest mean C-index and $AUC(t)$ across both states overall. This discriminative advantage was concentrated in the state following first disease progression, where prior transition history becomes available to the recurrent encoder, while performance at treatment initiation was comparable to the classical Cox-MSM. These results confirm that DG-MSM's architecture adapts across data regimes rather than requiring long trajectories to function.

Conditional simulation suggests that incorporating longitudinal visit information can improve long-term survival prediction relative to approaches that rely only on information available at the current treatment line. These findings indicate that the recurrent memory captures clinically meaningful aspects of disease evolution that are not represented by a single treatment state alone. Although the LoT-specific Cox benchmark model remained

competitive at the earliest treatment lines, DG-MSM showed increasing discriminative advantage from the third line onward, with prediction error also improving from the fourth line, particularly at later treatment lines where a more comprehensive treatment history is available. Taken together, these findings suggest that the predictive benefit of DG-MSM becomes more apparent as more longitudinal patient information becomes available.

The conditional simulation framework of DG-MSM provides a computational basis for VT approaches, in which a patient-specific generative model explores plausible future disease trajectories given an individual's observed history. Since the model's predictions update continuously as new clinical events are recorded without requiring re-training or a separately refit model at each treatment line, the framework follows the patients' treatment trajectory rather than providing a fixed forecast at a single time point. Instead of generating a single survival estimate, the model thus produces a distribution of possible future outcomes, allowing uncertainty in disease progression to be explicitly represented. For such a distribution to be clinically meaningful, the observed frequency of the outcome should correspond closely to the predicted probability among patients assigned that probability. We assessed this directly and found DG-MSM's simulations to be closely, though not perfectly, calibrated at treatment initiation, with prediction intervals mildly conservative rather than overconfident. Notably, calibration for OS, which requires the full recursive simulation to a terminal state, was only modestly worse than for sojourn time, which involves a single simulated transition, suggesting that any additional uncertainty introduced by repeated stochastic sampling remains limited in magnitude rather than compounding substantially across long simulated horizons. This property is a prerequisite specifically for VT applications, which require simulating arbitrarily far into a patient's future rather than over a single fixed horizon. As demonstrated in the dynamic prediction experiment, discriminative accuracy for residual OS also improved with increasing longitudinal follow-up, most notably from the third treatment line onward, suggesting that the model increasingly benefits from information accumulated throughout the patient's disease course. In principle, such calibrated simulations could inform clinically actionable questions, such as the probability of progression within a specific time horizon or the expected duration of the current treatment line.

While DG-MSM provides a joint framework to predict and simulate patient trajectories, several limitations should be acknowledged. The multi-state transition head operates on a discretized time scale, which introduces a resolution ceiling determined by the number of time bins. Although sufficient for modeling treatment-line transitions, this discretization may limit applicability to settings requiring finer temporal granularity. From a deployment perspective, patient-specific trajectory simulation relies on Monte Carlo sampling, requiring hundreds of forward passes through the recurrent architecture to obtain stable distributional estimates. While acceptable for retrospective research, this computational overhead may limit real-time integration into clinical workflows. Finally, although the architecture is designed to be disease-agnostic, our experiments focus exclusively on MM, and generalization to other domains with different state structures and transition dynamics remains to be demonstrated.

These limitations motivate several directions for future work. The discrete-time transition module could be replaced by continuous-time hazard models or neural ordinary differential equations (neural ODEs), removing the dependency on a fixed time grid and enabling application to settings with irregular or high-frequency observations [28,29]. Combining the generative trajectory model with counterfactual reasoning frameworks [30] would enable individualized treatment effect estimation under time-varying confounding. Finally, applying DG-MSM to other disease settings would validate the architecture's generalizability beyond oncology.

Methods

Model Description

DG-MSM uses variational inference to learn a generative model of the patient's covariates $(x_n)_{n=0}^{N}$ and trajectory $(v_n, \tau_n, e_n)_{n=1}^{N}$ as a history-aware jump process, where v_n is the sequence of visited states, τ_n is the time spent in the state, and e_n indicates a transition or censoring of the trajectory (Supplementary Fig. 9). We consider an equidistant discretization of the continuous time-scale [31] and define the calendar time of the process as $t_n := \sum_{m=1}^{n-1} \tau_m$. The model is structured into three main components: a variational

autoencoder for the baseline covariates, a recurrent variational module for the visit-specific covariates, and a discrete-time multi-state transition head that parametrises the distributions of events and waiting times **(Fig. 7)**.

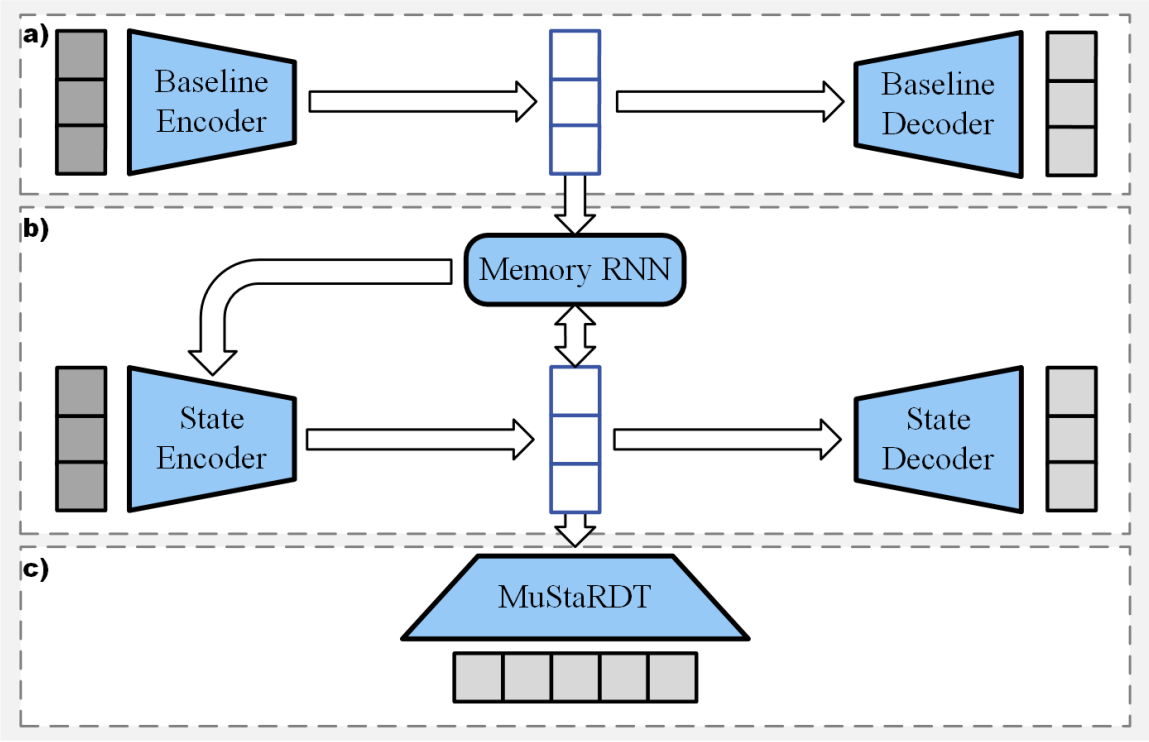


Fig. 7: Architecture of DG-MSM: **a)** Baseline HI-VAE, **b)** State-visit HI-VAE-recurrent neural network (RNN), **c)** MuStaRDT.

Baseline Covariates

First, we use a HI-VAE model [19] to learn the distribution of the baseline covariates x_0 while allowing missing values and mixed distribution types. This means introducing a latent representation z_0, a categorical mixture component s, and using the factorization $p(x_0, z_0, s) = p(x_0|z_0, s)p(z_0|s)p(s)$, where $p(z_0|s)p(s)$ is the mixture prior, and $p(x_0|z_0, s)$ is the covariate likelihood (shown in **Fig. 8a**). We factorize this joint likelihood over individual features, assigning each feature its own distribution family, e.g., Gaussian for real-valued, log-normal for positive, ordered logistic for ordinal, and categorical for discrete features. The recognition model of HI-VAE (**Fig. 8b**) uses a corresponding

approximate posterior $q(s, z_0|x_0) = q(s|x_0)q(z_0|s, x_0)$, which gives us the following evidence lower bound (ELBO) term

$$L_0 = E_{q(s,z_0|x_0)}[log\ p(x_0|z_0, s)] - E_{q(s|x_0)}[KL(q(z_0|x_0, s)||p(z_0|s))] - KL(q(s|x_0)||p(s)),$$

where the reconstruction term is evaluated only over observed features. The starting state v_1 can be either fixed or estimated, using an additional conditional distribution $p(v_1|z_0, s)$ contributing the cross-entropy loss $L_{init} = -E_{q(s,z_0|x_0)}[log\ p(v_1|z_0, s)]$ to the overall training objective (for the reported experiments, v_1 was fixed).

State-Visit Covariates

The estimation of state-visit covariates, events, and time points follows the sequential order of the process, starting at time $t_1 = 0$. Similarly to the baseline module, we introduce a visit-specific latent representation variable z_n for each visit. To allow current covariates to depend on the history of the whole process, we introduce a memory state h_{n-1}, implemented as the hidden state of an RNN [32]. This memory RNN is implemented as a gated recurrent unit (GRU) [33] and initialized with $h_0 = \phi_z(z_0)$, where ϕ_z is a learned feature transformation of the baseline latent representation into the dimension of the memory state. For each visit n and latent variable z_n, we define a prior $p(z_n|v_n, t_n, h_{n-1}, s)$ (**Fig. 8d**) and an approximate posterior $q(z_n|x_n, v_n, t_n, h_{n-1}, s)$ (**Fig. 8e**), where we condition on the mixture component s, current state v_n, calendar time t_n, and memory h_{n-1}. The visit-specific covariates are generated using a decoder for the feature-wise likelihood $p(x_n|z_n, h_{n-1}, s)$. Following the HI-VAE architecture, the decoder maps the latent to a per-feature intermediate representation before computing parameters for each feature's distribution. The memory state is then updated deterministically according to $h_n = RNN([\phi_v(v_n, t_n), \phi_z(z_n), h_{n-1}])$ (**Fig. 8c**), where ϕ_v maps the concatenation of state indicator and calendar time into the dimension of the memory state. h_n provides a low-dimensional summary of the entire trajectory up to and including visit n. Altogether, this gives us the corresponding ELBO contribution for the observed sequence $(x_n)_{n=1}^N$ of visit-specific covariates

$$L_N = \sum_{n=1}^{N} \; E_{q(s|x_0)q(\,z_n|\,x_n,v_n,t_n,h_{n-1},s)}[log\; p(x_n|z_n,h_{n-1},s)]$$
$$- E_{q(s|x_0)}[KL(\,q(z_n|x_n,v_n,t_n,h_{n-1},s)\;||\;p(z_n|v_n,t_n,h_{n-1},s)\,)],$$

where h_{n-1} is computed recurrently from the latent history $(z_m)_{m=0}^{n-1}$ and the past trajectory $(v_m,t_m)_{m=1}^{n-1}$.

Multi-State Modeling

To learn the distributions of events e_n and waiting times τ_n of the process, we consider the combined information $\psi_n := [s, z_0, z_n, h_{n-1}, t_n]$ (**Fig. 8f**). Including the time stamp t_n allows the transition model to capture non-stationary dynamics, such as hazards increasing with disease duration. Conditioned on the current state v_n, MuStaRDT models the next transition by factorizing the joint distribution $p(e_n,\tau_n|v_n,\psi_n) = p(e_n|v_n,\psi_n)p(\tau_n|e_n,v_n,\psi_n)$ and using a dedicated multi-layer perceptron (MLP) for each non-terminal state that first maps ψ_n to a state-specific intermediate representation to then compute $p(e_n|v_n,\psi_n)$ and $p(\tau_n|e_n,v_n,\psi_n)$ separately. The event distribution $p(e_n|v_n,\psi_n)$ is computed using a softmax function masked over admissible transitions from v_n. The probability distribution of event-specific waiting times is estimated using the sigmoid function to compute the discrete hazards $\lambda_n(k|e_n,v_n,\psi_n) = P[\tau_n = k|\tau_n > k-1, e_n, v_n, \psi_n]$ [31]. The corresponding probability mass function (PMF) is given by $p(\tau_n|e_n,v_n,\psi_n) = \lambda(\tau_n|e_n,v_n,\psi_n)S(\tau_n-1|e_n,v_n,\psi_n)$, where $S(\tau_n|e_n,v_n,\psi_n)$ is the event-specific survival function. This formulation yields a flexible discrete-time MSM with state-specific nonlinear transition dynamics and without parametric assumptions on the jump time distributions. The main loss contribution is given by the negative log-likelihood of a patient's observed trajectory $(v_n,\tau_n)_{n=1}^{N}$

$$L_{NLL} = \textstyle\sum_{n=1}^{N} \quad 1_{e_n\neq 0}(-\,log\; p(e_n|v_n,\psi_n) - log\; p(\tau_n|e_n,v_n,\psi_n)) - 1_{e_n=0}\; log\; (S(\tau_n|v_n,\psi_n)).$$

To improve calibration of survival estimates, we additionally use an auxiliary discrete ranked probability score on the marginal distribution, computed from the CDF of the waiting times [34]. The combined loss for the multi-state transitions is $L_{MuStaRDT} = L_{NLL} +$

γL_{score}, where $\gamma > 0$ is a hyperparameter scaling the scoring contribution. The final training objective combines the MuStaRDT loss with the ELBO contributions from the baseline HI-VAE L_0 and the trajectory RNN L_N using learned uncertainty weighting for multi-task learning [35]

$$L_{total} = e^{-\alpha_1} L_{MuStaRDT} + \alpha_1 - e^{-\alpha_2}\beta \ (L_N + L_0) + \alpha_2,$$

where α_1 and α_2 are learnable parameters, while $\beta > 0$ is a fixed hyperparameter. We optimize L_{total} using AdamW [36] with the learning rate reduced on a plateau. Continuous covariates are standardized using training-set statistics, which are stored and applied at test time. We apply early stopping based on the MuStaRDT loss contribution and restore the best model checkpoint. Full training hyperparameters and search spaces are provided in the supplementary material (Supplementary Table 3).

Altogether, we have a hierarchical generative model of the patient's observed trajectory and the latent dynamics

$$p(s, z_0, x_0, \{z_n, x_n, v_n, e_n, \tau_n\}_{n=1}^{N})$$
$$= p(s)p(z_0|s)p(x_0|z_0, s)p(v_1|z_0, s) \left[\prod_{n=1}^{N} p(z_n|v_n, t_n, h_{n-1}, s)p(x_n|z_n, h_{n-1}, s)p(e_n, \tau_n|v_n, \psi_n)\right],$$

where the time of state visit n is given by $t_n := \sum_{m=1}^{n-1} \tau_m$, and the next state v_{n+1} is uniquely determined by the event e_n. This combination of a variational recurrent encoder with a discrete-time MSM yields transition probabilities informed by the full patient history through learned recurrent memory, without imposing parametric hazard assumptions.

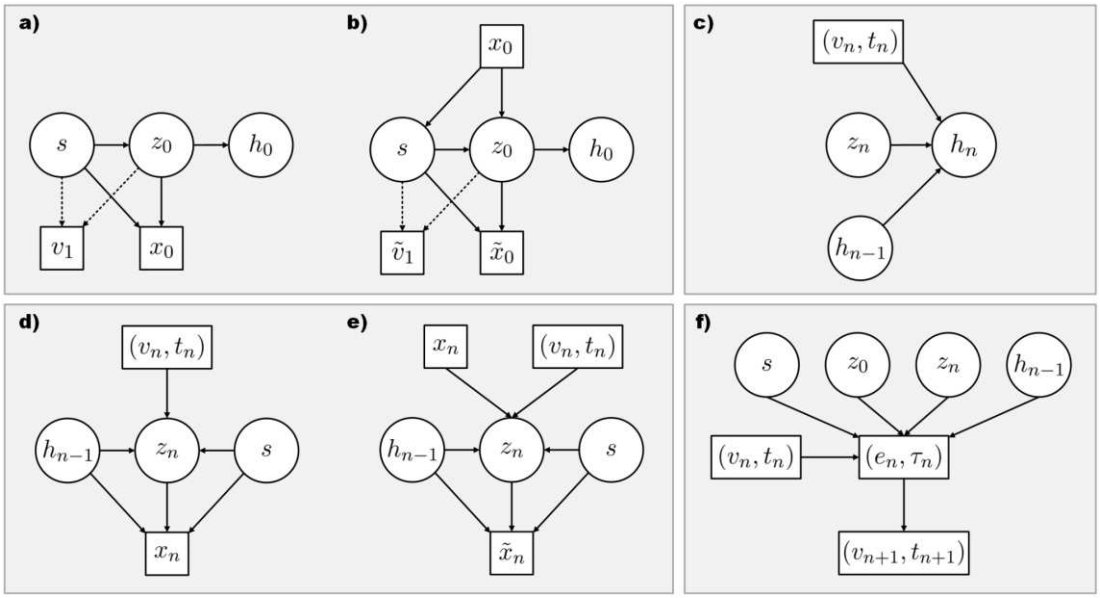


Fig. 8: Computational paths of DG-MSM. **a)** Generative model for baseline covariates and the initial state, **b)** Recognition model for baseline covariates and the initial state, **c)** Memory state update, **d)** Generative model for state-visit covariates, **e)** Recognition model for state-visit covariates, and **f)** MuStaRDT estimation.

Simulating Patient Trajectories

DG-MSM can be used to simulate future trajectories conditioned on an individual patient's observed history. Given baseline covariates and a trajectory up to visit m, we use the approximate posterior to obtain a patient-specific latent representation and memory state summarising the history up to t_m. This history can be encoded either deterministically (using posterior means) or stochastically (sampling from the approximate posterior for each path). We then simulate the future from visit $m+1$ onward by repeatedly sampling visit-level latents, covariates, and transitions. For $n = m+1, m+2 \ldots$:

1. A visit-level latent variable is drawn: $z_n \sim p(z_n | v_n, t_n, h_{n-1}, s)$.
2. Visit covariates are sampled: $x_n \sim p(x_n | z_n, h_{n-1}, s)$.
3. The transition type and time are sampled: $e_n \sim p(e_n | v_n, \psi_n)$ and $\tau_n \sim p(\tau_n \mid e_n, v_n, \psi_n)$.
4. The memory state is updated: $h_n = RNN([\phi_v(v_n, t_n), \phi_z(z_n), h_{n-1}])$.

5. The state and calendar time are updated: $t_{n+1} = t_n + \tau_n$ and v_{n+1} is the state after transition e_n.

The process stops when a terminal state is reached or is censored when a maximum path length is exceeded, producing a set of simulated multi-state trajectories and visit-level covariates. A path may also be right-censored earlier if the sampled waiting time for a given visit falls outside the model's discrete time horizon. Since the full joint distribution is intractable due to the sequential dependence of latent variables through the recurrent memory, these forward simulations provide Monte Carlo estimates of predictive quantities such as state-occupation probabilities or Kaplan-Meier curves under the fitted model.

We assess the model's ability to predict events dynamically from different points in a patient's disease course. For each non-terminal state visited by a patient in the test set, we condition the model on the observed history up to and including entry into that state, and mask all future visits from the observed data. Subsequently, we generate multiple independent future paths using the forward simulation procedure. From the generated trajectories, we estimate residual survival using a Kaplan-Meier estimator of the simulated durations and event indicators. The corresponding ground-truth outcome is each patient's observed residual OS.

Calibration and Evaluation Metrics

We base our evaluation on the metrics proposed by Kvamme et al. and implemented in the pycox package[15]. As a measure of discrimination, we use the time-dependent C-index of Antolini et al.[23], computed from the predicted state-specific survival curves and the observed sojourn times. In addition, we consider an adaptation of this time-dependent C-index, following the inverse censoring weighting approach of Uno et al.[37], where each concordant pair's contribution is scaled by IPCW computed using the KM estimator of the censoring times. This yields a censoring-adjusted concordance measure that is particularly suitable for the often heavily censored sojourn times in later disease or treatment states. To assess overall prediction error, we compute the time-dependent Brier score corrected for censoring using IPCW. It measures the squared deviation between

the observed event indicator and the predicted survival probability at a given time point. We summarise this as the IBS, obtained by numerically integrating the time-dependent Brier score over the evaluation time grid[38]. Additionally, we consider a time-dependent AUC approach using IPCW weighting[39]. The $AUC(t)$ quantifies how well the model discriminates between individuals who have experienced a transition by time t and those who remain in the current state at time t.

To assess the reliability of DG-MSM's simulated predictive distributions, we use D-calibration[24], which tests whether a patient's observed outcome is consistent with their individually predicted distribution. For each patient, we compute the probability-integral transform (PIT) of their observed outcome under their model-predicted cumulative distribution function. Under correct calibration, this quantity is uniformly distributed across the population. We partition the unit interval into 10 equal-width bins and compare observed against expected bin counts using a χ^2 goodness-of-fit test. We additionally assess prediction-interval coverage using the prediction interval coverage probability ($PICP(\alpha)$[25]) computed across a range of nominal confidence levels α. For each α, we construct a corresponding prediction interval from each patient's simulated outcome distribution and compute the proportion of patients whose observed outcome falls within this interval. Because observed outcomes are subject to right censoring, we compute $PICP(\alpha)$ using IPCW weights, restricted to patients with an observed event. $PICP(\alpha)$ values exceeding the nominal level α indicate prediction intervals that are conservative rather than overconfident.

Data

To evaluate the performance of the investigated models across different data types, we designed two use cases: one leveraging real-world data and the other leveraging clinical data. We used real-world data from the TriNetX Oncology network and clinical data from the CoMMpass study. Patients from the TriNetX database were included if they were diagnosed with MM after January 1, 2010, and had received at least one LoT. Within each dataset, a set of features was selected as baseline or state-dependent covariates. The

selection of covariates was dataset-specific, based on the available measurements and the underlying data structure. In the CoMMpass dataset, six features (e.g., age, sex, and ISS^3) were selected as baseline covariates and seven laboratory measurements as state-dependent covariates, including the serum M protein value. In contrast, in the TriNetX dataset, 12 baseline covariates, including comorbidities, and eight state-specific covariates were selected for all LoTs. The cohort-specific covariates are presented in supplementary Table 4. Following data filtering and accounting for the available covariates, we further restricted the analysis to patients with complete observations across all selected covariates. While our proposed model is capable of handling missing data, this restriction was necessary to enable a fair comparison with benchmark models that cannot account for incomplete inputs. As a result, 1,008 patients in the CoMMpass dataset and 8,878 in the TriNetX dataset were included in each experiment. More details on the patient selection can be found in the supplementary material (Supplementary Fig. 10-15).

References

1. Schavgoulidze, A., Cazaubiel, T., Perrot, A., Avet-Loiseau, H. & Corre, J. Multiple Myeloma: Heterogeneous in Every Way. *Cancers* **13**, 1285 (2021).

2. Rajkumar, S. V. Multiple myeloma: 2024 update on diagnosis, risk-stratification, and management. *Am. J. Hematol.* **99**, 1802–1824 (2024).

3. Greipp, P. R. *et al.* International Staging System for Multiple Myeloma. *J. Clin. Oncol.* **23**, 3412–3420 (2005).

4. Palumbo, A. *et al.* Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group. *J. Clin. Oncol.* **33**, 2863–2869 (2015).

5. D'Agostino, M. *et al.* Second Revision of the International Staging System (R2-ISS) for Overall Survival in Multiple Myeloma: A European Myeloma Network (EMN) Report Within the HARMONY Project. *J. Clin. Oncol.* **40**, 3406–3418 (2022).

6. Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B. & Bender, A. Deep learning for survival analysis: a review. *Artif. Intell. Rev.* **57**, 65 (2024).

7. Maura, F. *et al.* Genomic Classification and Individualized Prognosis in Multiple Myeloma. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **42**, 1229–1240 (2024).

8. Bhatt, P., Kloock, C. & Comenzo, R. Relapsed/Refractory Multiple Myeloma: A Review of Available Therapies and Clinical Scenarios Encountered in Myeloma Relapse. *Curr. Oncol. Tor. Ont* **30**, 2322–2347 (2023).

9. Parekh, D. S. *et al.* Updates on Therapeutic Strategies in the Treatment of Relapsed/Refractory Multiple Myeloma. *Cancers* **16**, 2931 (2024).

10. Katsoulakis, E. *et al.* Digital twins for health: a scoping review. *Npj Digit. Med.* **7**, 77 (2024).

11. Tudor, B. H. *et al.* A scoping review of human digital twins in healthcare applications and usage patterns. *Npj Digit. Med.* **8**, 587 (2025).

12. Weirauch, U. *et al.* Design specifications for biomedical virtual twins in engineered adoptive cellular immunotherapies. *Npj Digit. Med.* **8**, 493 (2025).

13. Cid Ruzafa, J. *et al.* Patient population with multiple myeloma and transitions across different lines of therapy in the USA: an epidemiologic model: Multiple Myeloma Lines of Therapy. *Pharmacoepidemiol. Drug Saf.* **25**, 871–879 (2016).

14. Lee, C., Zame, W., Yoon, J. & Schaar, M. van der. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proc. AAAI Conf. Artif. Intell.* **32**, (2018).

15. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-Event Prediction with Neural Networks and Cox Regression. Preprint at https://doi.org/10.48550/arXiv.1907.00825 (2019).

16. Apellániz, P. A., Parras, J. & Zazo, S. CR-SAVAE: A Parametric Method for Survival Analysis with Competing Risks. in *2024 32nd European Signal Processing Conference (EUSIPCO)* 1526–1530 (IEEE, Lyon, France, 2024). doi:10.23919/EUSIPCO63174.2024.10715431.

17. De Wreede, L. C., Fiocco, M. & Putter, H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Comput. Methods Programs Biomed.* **99**, 261–274 (2010).

18. Andersen, P. K. & Keiding, N. Multi-state models for event history analysis. *Stat. Methods Med. Res.* **11**, 91–115 (2002).

19. Nazábal, A., Olmos, P. M., Ghahramani, Z. & Valera, I. Handling incomplete heterogeneous data using VAEs. *Pattern Recognit.* **107**, 107501 (2020).

20. Wendland, P. *et al.* Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations. *Npj Digit. Med.* **5**, 122 (2022).

21. Makarov, N. *et al.* Large language models forecast patient health trajectories enabling digital twins. *Npj Digit. Med.* **8**, 588 (2025).

22. Lammert, J. *et al.* Large language models-enabled digital twins for precision medicine in rare gynecological tumors. *Npj Digit. Med.* **8**, 420 (2025).

23. Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Stat. Med.* **24**, 3927–3944 (2005).

24. Haider, H., Hoehn, B., Davis, S. & Greiner, R. Effective Ways to Build and Evaluate Individual Survival Distributions. *J. Mach. Learn. Res.* **21**, 1–63 (2020).

25. Khosravi, A., Nahavandi, S., Creighton, D. & Atiya, A. F. Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. *IEEE Trans. Neural Netw.* **22**, 1341–1356 (2011).

26. Goldstein, M., Han, X., Puli, A., Perotte, A. & Ranganath, R. X-CAL: Explicit Calibration for Survival Analysis. in *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 18296–18307 (Curran Associates, Inc., 2020).

27. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Routledge, 2013). doi:10.4324/9780203771587.

28. Tang, W., Ma, J., Mei, Q. & Zhu, J. SODEN: A Scalable Continuous-Time Survival Model through Ordinary Differential Equation Networks. *J. Mach. Learn. Res.* **23**, 1–29 (2022).

29. Groha, S., Schmon, S. M. & Gusev, A. A General Framework for Survival Analysis and Multi-State Modelling. Preprint at https://doi.org/10.48550/ARXIV.2006.04893 (2020).

30. Bica, I., Alaa, A. M., Jordon, J. & van der Schaar, M. Estimating Counterfactual Treatment Outcomes over Time Through Adversarially Balanced Representations. https://doi.org/10.48550/ARXIV.2002.04083 (2020) doi:10.48550/ARXIV.2002.04083.

31. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.* **27**, 710–736 (2021).

32. Chung, J. *et al.* A Recurrent Latent Variable Model for Sequential Data. in *Advances in Neural Information Processing Systems* (eds Cortes,

C., Lawrence, N., Lee, D., Sugiyama, M. & Garnett, R.) vol. 28 (Curran Associates, Inc., 2015).

33. Cho, K. *et al.* Learning phrase representations using RNN encoder–decoder for statistical machine translation. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* 1724–1734 (2014).

34. Avati, A. *et al.* Countdown Regression: Sharp and Calibrated Survival Predictions. in *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference* (eds Adams, R. P. & Gogate, V.) vol. 115 145–155 (PMLR, 2020).

35. Kendall, A., Gal, Y. & Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. Preprint at https://doi.org/10.48550/ARXIV.1705.07115 (2017).

36. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at https://doi.org/10.48550/ARXIV.1711.05101 (2017).

37. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* **30**, 1105–1117 (2011).

38. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**, 2529–2545 (1999).

39. Blanche, P., Dartigues, J. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat. Med.* **32**, 5381–5397 (2013).

Author Contributions

Holger Fröhlich designed the study and acquired funding. Sebastian Schwick, Shammi More, and Guillaume Azarias prepared the datasets. Sebastian Schwick conducted the analyses and generated the results. Yasamin Salimi contributed to the analyses. Sebastian Schwick wrote the original manuscript draft. Yasamin Salimi, Shammi More, Guillaume Azarias, and Holger Fröhlich reviewed and revised the manuscript.

Acknowledgements

The CERTAINTY project is funded by the European Union (Grant Agreement 101136379). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

These data were generated as part of the Multiple Myeloma Research Foundation Personalized Medicine Initiatives (https://research.themmrf.org and www.themmrf.org).

Competing interests

The authors declare no competing interests related to this work.

Data availability

CoMMpass data is available from the respective data owners upon request: https://themmrf.org/.

Code availability

The model code is available on GitHub (https://github.com/SCAI-BIO/dg-msm).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DGMSMsupp.docx