Synthesizing multidimensional clinical profiles from published Kaplan–Meier images

Zheqing Zhu[1,2], Fangfang Shen[3], Yuhua Qian[1,2,*], and Jieting Wang[1,2]

[1]Institute of Big Data Science and Industry, Shanxi University, Taiyuan, Shanxi, China
[2]Key Laboratory of Evolutionary Science Intelligence of Shanxi Province, Taiyuan, Shanxi, China
[3]Shanxi Hospital Affiliated to Cancer Hospital, Chinese Academy of Medical Sciences, Taiyuan, Shanxi, China

ABSTRACT

Clinical decision-making relies on understanding intersectional treatment effects across multiple patient characteristics. However, randomized controlled trials typically report one-dimensional marginal summaries, obscuring the underlying joint distributions of these characteristics. To address this, we developed MD-JoPiGo, a computational framework that reconstructs multidimensional clinical profiles from published 1D Kaplan–Meier curves. The approach utilizes the maximum entropy principle to estimate joint stratum frequencies and applies simulated annealing to generate individual-level data. We show that reconstruction fidelity depends on the underlying causal topology. Parallel predictors are resolved unconditionally, whereas interdependent structures require minimal structural priors to resolve unidentifiability. In evaluations using simulated data and empirical cohorts (lung cancer, $n = 228$; colon cancer, $N = 929$), the framework accurately recovered unobserved multivariable dynamics. Applied to fragmented and temporally misaligned reports from the CheckMate 227 trial, MD-JoPiGo reconstructed latent intersectional efficacy consistent with the clinical ground truth. By synthesizing multivariable evidence from 1D margins, this framework enables the secondary analysis of historical RCTs, supporting IPD meta-analyses and synthetic trial emulations.

Introduction

Precise therapeutic decision-making in clinical medicine requires an understanding of how treatment effects vary across intersecting clinical dimensions[1]. Precision medicine increasingly moves beyond population-level average effects to identify specific multivariable subcohorts that derive maximal clinical benefit[2]. Indeed, multivariable predictive modeling has revealed clinically important heterogeneity of treatment effects (HTE) in a substantial proportion of randomized trials, demonstrating that average results often mask critical subcohort variations[3]. For instance, determining the optimal response to immune checkpoint inhibitors relies heavily on the multivariable interplay between biomarker expression and specific histological subtypes[4]. However, the multivariable clinical profiles necessary to resolve these interactions are effectively locked within proprietary trial silos, as strict privacy regulations and commercial confidentiality frequently prohibit the sharing of individual patient-level data[5,6]. This data isolation forces the scientific community to rely on fragmented summary data; even massive-scale meta-analyses encompassing millions of patients are typically restricted to reporting isolated marginal effects, leaving the underlying joint distributions of patient characteristics fundamentally unobservable[7].

The primary barrier to observing these joint distributions lies in the dimensionality reduction inherent to published trial reports. To accommodate reporting brevity and space constraints, high-dimensional patient profiles are routinely compressed into one-dimensional (1D) summaries, predominantly presented as Kaplan–Meier (KM) curves[8–10]. Although the intersection of certain clinical variables could provide critical insights for tailored therapeutic decisions[11], these interactions are largely obscured within 1D plots. Consequently, researchers are restricted from evaluating these nuanced relationships, precluding the in-depth, multivariable analyses necessary to identify which specific patient subgroups derive the greatest benefit[12]. Failing to account for these within-trial multivariable interactions risks falling into the "daft" or "deluded" analytical traps described by Fisher et al., where ecological bias from aggregate summaries can lead to misleading clinical conclusions[13].

Current methodologies for IPD reconstruction facilitate the extraction of one-dimensional survival trajectories from published KM curve images[14,15]. Although these tools can reconstruct isolated treatment arms or specific subgroups, they operate strictly within a single dimension. They lack the statistical framework to synthesize independent 1D marginals into a coherent joint distribution. Without a mechanism to estimate the intersectional frequencies across these marginal constraints, the multivariable profiles of the original trial remain fragmented and analytically inaccessible[16].

To transition from isolated 1D reconstructions to multivariable trial emulation, we introduce MD-JoPiGo (Multidimensional Joint Patient Individual-data Generator and Optimizer). This framework synthesizes multidimensional clinical profiles by integrating one-dimensional individual patient data (1D-IPD) extracted from published Kaplan–Meier images using the web-based KM-PoPiGo tool (`https://kmpopigo.github.io`). The methodology employs a two-stage reconstruction: first,

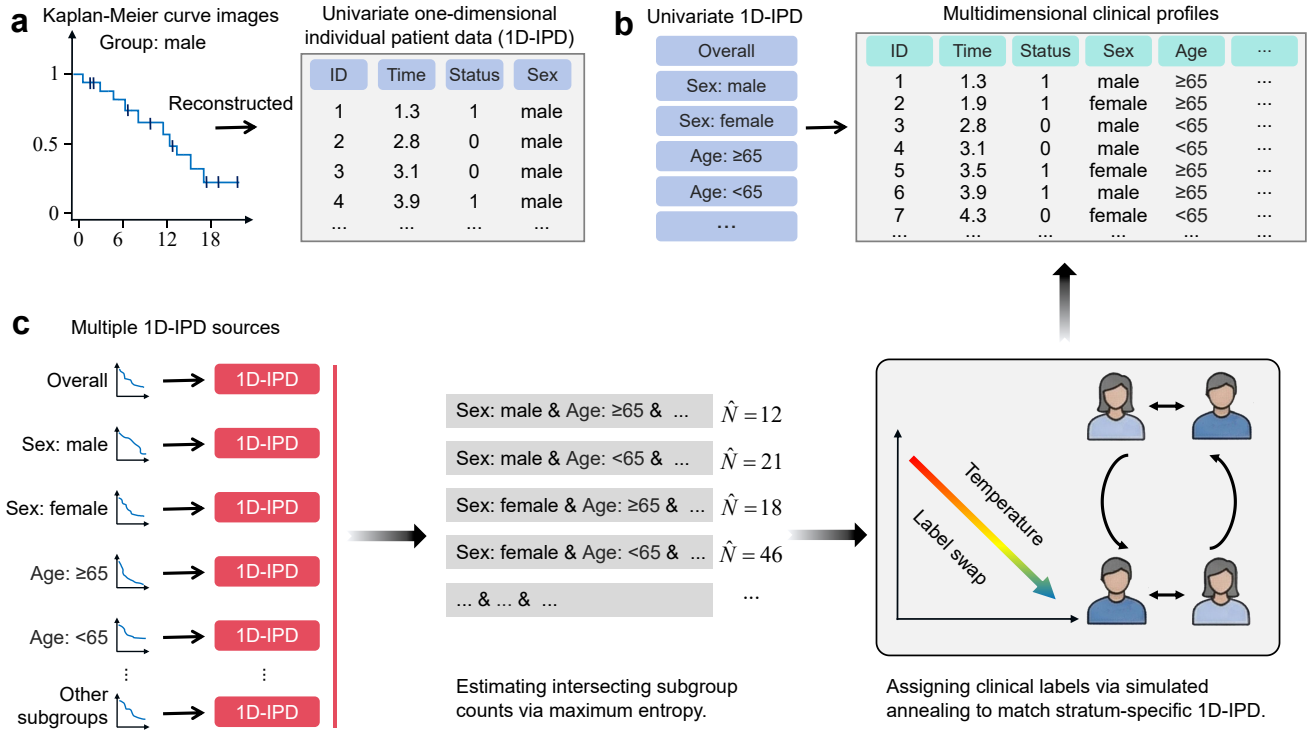

Fig. 1 | Schematic of the MD-JoPiGo framework for synthesizing multidimensional clinical profiles from published Kaplan–Meier curve images. a. Reconstruction of one-dimensional individual patient data (1D-IPD). Published Kaplan–Meier (KM) curve images (e.g., the male subgroup) are digitized to extract individual survival times and event statuses. We define this extracted data as 1D-IPD, given that it maps survival outcomes to only one clinical dimension at a time. **b. Transition from fragmented 1D-IPD to a unified multidimensional cohort.** Separate 1D-IPD sets (e.g., overall, sex-, and age-stratified populations) are integrated to synthesize a single, comprehensive cohort. In this unified dataset, each patient is assigned a complete set of concurrent clinical labels. **c. The two-stage optimization pipeline.** First, a maximum entropy approach estimates the unobserved joint subgroup frequencies ($\hat{N}$; e.g., female patients aged < 65) from the available 1D-IPD. Second, simulated annealing assigns these combinatorial labels to individual patients. A temperature-controlled, iterative label-swapping mechanism optimizes this distribution until the synthesized multidimensional profiles accurately reproduce the original stratum-specific survival trajectories.

a maximum entropy approach estimates joint frequencies across clinical strata; subsequently, a simulated annealing-driven label exchange generates individual-level profiles that honor the observed survival dynamics. These unified outputs function as synthetic cohorts, providing a quantitative foundation for synthetic control arm (SCA) construction and the estimation of intersectional treatment effects.

We find that synthesis fidelity depends on the causal topology of the clinical covariates. While parallel prognostic factors (e.g., PD-L1 expression and tumor mutational burden) are accurately resolved via unconstrained optimization, complex topologies involving chain mediation (e.g., age-related frailty) or collider structures induce systematic biases, such as localized coefficient drift or artificial negative correlations. MD-JoPiGo corrects these topological distortions by integrating minimal structural priors—specifically, a single intersectional proportion—ensuring the structural identifiability of the joint distribution across diverse causal scenarios.

We validated MD-JoPiGo through a hierarchical approach encompassing systematic simulations, complete empirical datasets (a lung cancer ground truth cohort, $n = 228$[17]; a stage III colon cancer RCT, $N = 929$[18]), and fragmented, asynchronous reports from the CheckMate 227 trial. These validations confirm the theoretical structural boundaries: parallel predictors (as demonstrated in the colon cancer cohort) are reliably recovered without structural constraints, whereas resolving interdependent variables (such as age and performance status in the lung cancer cohort) and simulation-validated collider structures requires minimal structural priors to accurately recover the true joint distribution. Applied to the CheckMate 227 trial, the framework successfully reconstructed latent intersectional efficacy from temporally misaligned 1D margins. By synthesizing multivariable individual-level data from published trial summaries, this approach provides a scalable quantitative foundation for cross-study

benchmarking and synthetic trial emulation.

Results

Performance and adaptation of MD-JoPiGo in parallel, mediated, and collider causal scenarios

Transitioning 1D individual patient data (1D-IPD) into multidimensional profiles requires accurate causal mapping. Without observed joint distributions, MD-JoPiGo defaults to a maximum entropy (MaxEnt) assumption, treating predictors as conditionally independent. To define the boundary conditions of this baseline assumption, we evaluated three fundamental clinical topologies using directed acyclic graphs (DAGs)[19]: (1) parallel independence (Scenario A), where predictors such as tumor genotype and geographic region affect survival independently; (2) chain mediation (Scenario B), where one predictor influences another (e.g., advanced age (X_1) contributing to clinical frailty (X_2)); and (3) collider selection (Scenario C), where independent variables jointly determine trial enrollment (C). In the latter scenario, high disease severity and high socioeconomic access may both increase admission likelihood, inducing an artificial negative correlation among enrolled patients (Berkson's paradox). Using Monte Carlo simulations ($n = 200$, $N = 200$ per cohort) across these topologies (Figure 2), we benchmarked ground-truth multivariable hazard ratios against two MD-JoPiGo modes: the uncalibrated mode (pure MaxEnt) and the structurally calibrated mode.

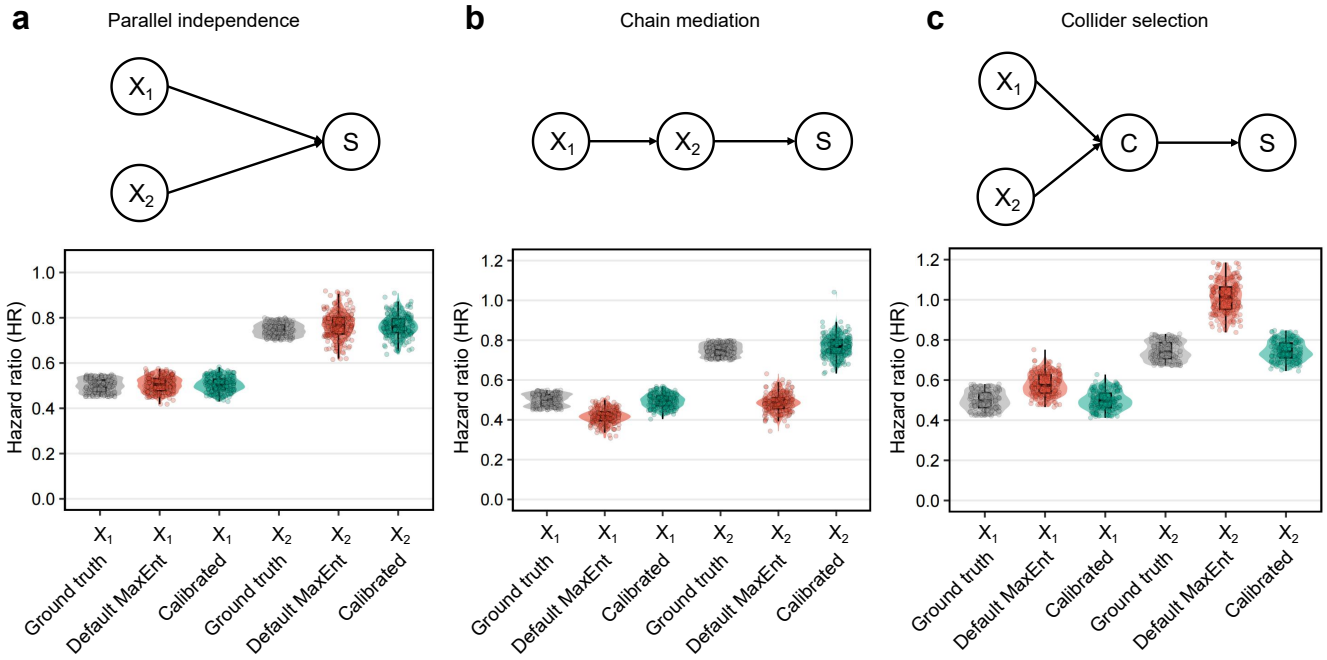

Fig. 2 | Reconstruction of multivariable hazard ratios across distinct causal topologies using MD-JoPiGo. Monte Carlo simulations ($n = 200$ iterations, $N = 200$ per cohort) were conducted. All presented hazard ratios (HRs) are multivariable-adjusted, estimated via Cox proportional hazards models incorporating both X_1 and X_2. In the directed acyclic graphs (DAGs), S denotes survival outcome, and C denotes a collider variable representing selection criteria. **(a) Parallel independence:** Under strict predictor independence, the default MaxEnt reconstruction (red) aligns with the ground truth distribution (gray), demonstrating that the default assumption of conditional independence is sufficient when there is no inter-variable correlation. **(b) Chain mediation:** When X_1 sequentially influences X_2, the default MaxEnt approach (red)—which relies on the baseline assumption of conditional independence—results in coefficient drift towards the null. The calibrated mode (green), incorporating empirical joint probability priors, corrects this bias. **(c) Collider selection:** Conditioning on C induces negative correlation between otherwise independent predictors. The default MaxEnt reconstruction (red) fails to account for this induced correlation, whereas the calibrated approach (green) recovers the ground truth distribution. Box plots indicate median and interquartile range (IQR); violins display the density distribution of HR estimates.

Under the baseline condition of parallel independence, where the mathematical assumption of maximum entropy matches causal reality, the default MaxEnt reconstruction yielded multivariable HRs that closely aligned with the ground truth distribution. However, clinical variables are rarely strictly independent. In chain mediation, the default MaxEnt approach—ignoring the sequential dependency between predictors—imposed an artificial structure of conditional independence upon the cohort. This

structural mismatch manifested as coefficient drift, systematically attenuating the reconstructed HRs toward the null. By introducing the calibration module of MD-JoPiGo, which ingests empirical joint probability priors (e.g., cross-tabulation counts from baseline characteristics) to override the MaxEnt assumption, the algorithm corrected this bias and recovered the ground truth HRs.

Furthermore, under the condition of collider selection, the default MaxEnt reconstruction failed to account for the trial-specific selection bias, resulting in distorted multivariable risk estimates. In contrast, the calibrated mode anchored the reconstruction to the observed joint distribution of the selected cohort. This mechanism demonstrated that while MD-JoPiGo does not inherently "de-bias" the original trial's inclusion criteria, it ensures fidelity to the specific multivariable dynamics of the selected population, preventing secondary topological distortion during the IPD synthesis process. Collectively, these simulations demonstrate the necessity of structural calibration, validating MD-JoPiGo as a robust framework capable of preserving multivariable fidelity across diverse causal topologies.

Synthesis and structural calibration of multidimensional profiles in chain mediation topologies

Single-arm clinical trials frequently rely on synthetic control arms to evaluate relative treatment efficacy. Constructing these controls, however, requires multidimensional clinical profiles to adjust for baseline confounding. Because published reports typically present high-dimensional cohort data as independent one-dimensional (1D) Kaplan-Meier curves, the underlying joint distributions of patient characteristics remain unobservable. To determine whether MD-JoPiGo can recover these profiles from isolated marginal summaries, we applied the framework to a reference lung cancer dataset ($n = 228$)[17], emulating a published single-arm cohort. This dataset is characterized by a chain mediation topology, wherein advanced age influences overall survival through the mediating variable of ECOG performance status.

The default maximum entropy (MaxEnt) method of the MD-JoPiGo framework synthesized multidimensional clinical profiles from 1D-IPD sets reconstructed from overall and stratified Kaplan-Meier curves. These unconstrained synthetic trajectories were statistically concordant with the reference ground truth at the one-dimensional level (Female; log-rank $P = 0.95$, Fig. 3a). To ensure representativeness, the displayed examples reflect subcohorts with median log-rank P-values within their respective tiers. However, this concordance exhibited a dimensional gradient, declining across two-dimensional interactions (Male and ECOG PS 0-1; $P = 0.85$) and reaching a statistical boundary in sparse three-dimensional combinations (Male, Age < 65, and ECOG PS 0-1; $P = 0.65$). This attenuation of statistical alignment is quantified by the uncalibrated log-rank results (Fig. 3b, empty grey circles). Survival comparisons for all remaining subcohorts are provided in Supplementary Fig. 1.

To evaluate internal covariance within the synthesized profiles, we analyzed independent prognostic weights via multivariable Cox regression and risk prediction using clinical nomograms. While the adjusted hazard ratios (HR) for sex and ECOG performance status matched the reference data, unconstrained synthesis overestimated the prognostic effect of age (≥ 65 vs < 65). Because the default MaxEnt approach assumes conditional independence, it omits the inherent collinearity between advanced age and poorer performance status. This assumption redistributed shared prognostic variance, yielding an inflated HR for age (1.75 vs 1.18) and spurious statistical significance ($P = 0.003$ vs reference $P = 0.395$; Fig. 3c, cyan triangles). In the derived multivariable nomograms, age was correspondingly allocated disproportionate points, although the overall risk factor hierarchy (ECOG PS $>$ Sex $>$ Age) and survival prediction scales remained comparable to the reference models (Supplementary Fig. 2).

We recalibrated the synthesis using a minimal structural prior—the cross-tabulated proportions of age and ECOG performance status—to constrain the underlying joint distribution and correct the localized coefficient drift. This recalibration systematically improved survival concordance across unobserved two- and three-dimensional interactions, elevating the minimum subcohort log-rank alignment from $P = 0.05$ to $P = 0.32$ (Fig. 3b, solid blue dots; Supplementary Fig. 2). Within the multivariable Cox regression, the structural calibration resolved the internal covariance deviation: the adjusted HR for age (1.20; $P = 0.351$) reverted to the non-significant baseline, and the primary prognostic effects of sex and performance status remained consistent with the ground truth (Fig. 3c, red circles). To evaluate this topological sensitivity, we assessed all possible calibration combinations (Supplementary Table 1). The analysis demonstrated that synthesis fidelity relies on calibrating the specific causal dependency. Providing the functionally distinct Sex-ECOG prior did not correct the age-related bias (HR 1.64). In contrast, the minimal Age-ECOG prior eliminated the coefficient drift, achieving structural identifiability comparable to a fully calibrated topology. Minimal structural anchors thus counteract the variance redistribution inherent to unconstrained maximum entropy, generating structurally consistent joint distributions for precise trial emulation.

Intersectional treatment effect estimation in a simulated clinical trial cohort

Identifying treatment effect heterogeneity across distinct patient subgroups is central to precision oncology[12]. While the joint distribution of baseline covariates determines specific treatment benefits, published randomized controlled trials (RCTs) typically collapse these dimensions into isolated marginal stratifications. This "dimensionality gap" often precludes the

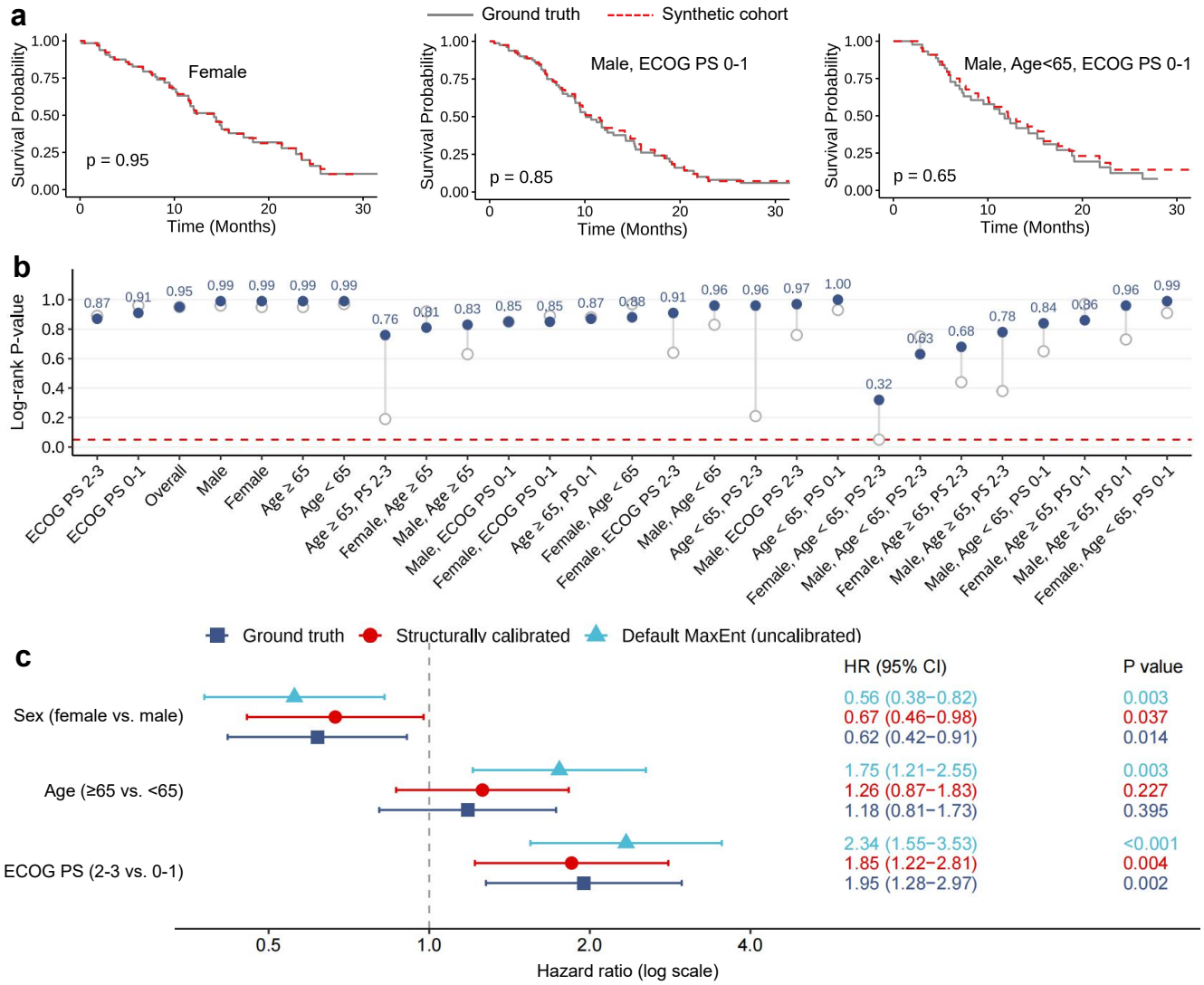

Fig. 3 | Structural calibration of synthesized multidimensional clinical profiles in an empirical lung cancer cohort **a**,Kaplan–Meier survival trajectories of representative one-, two-, and three-dimensional subcohorts synthesized under unconstrained default maximum entropy (MaxEnt) assumptions. Displayed panels represent subcohorts with median log-rank *P*-values within each dimensional tier. **b**, Global assessment of subcohort survival concordance. Empty grey circles represent the uncalibrated synthesis, showing the attenuation of statistical alignment as combinatorial sparsity increases. Solid blue dots show the systematic recovery of survival concordance across unobserved interactions after incorporating a minimal structural prior (cross-tabulated proportions of age and ECOG performance status). The red dashed line marks the $P = 0.05$ significance threshold. **c**, Multivariable Cox regression analysis of independent prognostic weights. Uncalibrated synthesis (cyan triangles) exhibits localized coefficient drift, inflating the hazard ratio for age. Structural calibration (red circles) corrects this internal covariance deviation, restoring adjusted hazard ratios to the reference ground truth (dark blue squares). Error bars indicate 95% confidence intervals.

analysis of intersectional effects. Consequently, emulating trials through synthetic cohorts necessitates the reconstruction of high-dimensional profiles directly from these aggregated marginal summaries.

To assess the real-world robustness of MD-JoPiGo, we utilized a randomized controlled trial (RCT) dataset of stage III colon cancer ($N = 929$) comparing adjuvant levamisole plus 5-fluorouracil (Lev+5FU) versus observation[18]. To replicate the restricted data environments of published literature, we generated paired Kaplan–Meier plots from the reference IPD for the overall population and 1D-stratified subgroups defined by node status, age, and sex. These prognostic factors exhibit a topology of parallel independence (Scenario A), as node status, age, and sex operate as largely autonomous predictors of survival in this

clinical context. These graphical summaries were digitized via KM-PoPiGo to reconstruct the individual patient data (IPD). Reflecting typical reporting limitations, the synthesis was constrained exclusively by sample sizes, at-risk tables, and censoring times, purposefully omitting specific event counts.

Treatment efficacy was then evaluated within unobserved 2D interaction subgroups, representing a clinically significant granularity for decision-making. Across these subcohorts, synthesized treatment hazard ratios (HR) remained consistent with the reference ground truth, with overlapping 95% confidence intervals (Fig. 5a). The framework recovered prognostic dynamics even in highly stratified subsets; for example, the survival trajectories and prognostic HR for node status among Lev+5FU-treated female patients aligned with actual data (Fig. 5b). This structural concordance was consistently observed across broader multidimensional stratifications defined by either demographic or clinical baseline characteristics (Supplementary Figs. 4 and 5). Multivariable Cox regression confirmed that the independent treatment effect and baseline covariate weights were preserved without significant coefficient drift (Fig. 5c), reflecting the lower collinearity among the trial variables compared to frailty-related indicators. MD-JoPiGo thus derives intersectional treatment effects and multi-variable trial dynamics directly from standard marginal inputs.

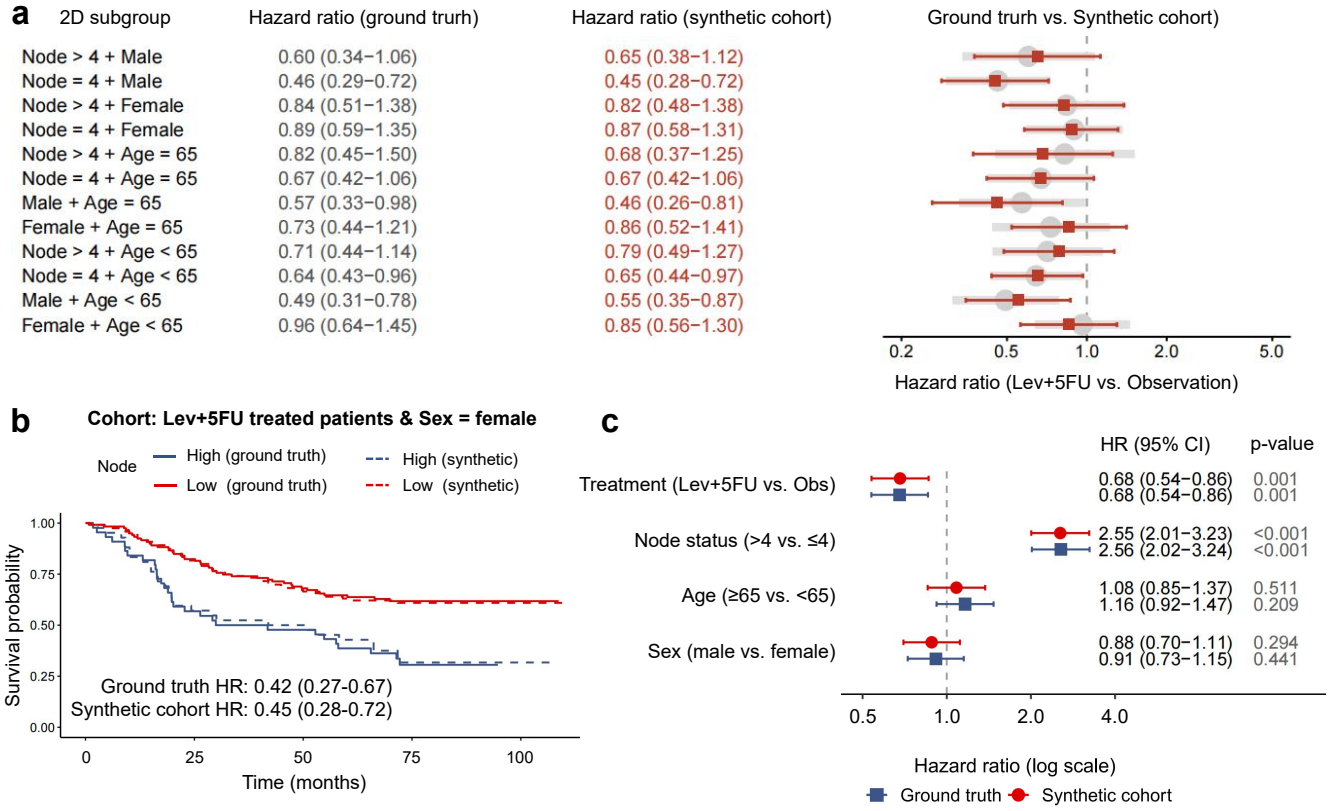

Fig. 4 | Intersectional treatment effect estimation in a synthesized clinical trial cohort. a, Forest plot evaluating the treatment effect (Hazard Ratio of Lev+5FU vs. Observation) across unobserved 2D interaction subgroups. The synthetic cohort (red squares) maintains statistical alignment with the reference ground truth (grey circles). **b**, Kaplan–Meier survival trajectories and corresponding prognostic HR for a granular 3D subcohort (Lev+5FU-treated female patients), comparing Node > 4 (blue) versus Node ≤ 4 (red). Solid lines denote the ground truth, while dashed lines represent the synthesized profiles. **c**, Multivariable Cox proportional hazards regression for the entire simulated trial cohort, verifying the independent prognostic weights of the assigned treatment and baseline covariates. Horizontal error bars indicate 95% CI.

Reconstructing latent intersectional efficacy from fragmented and asynchronous trial reports

Although previous sections validated MD-JoPiGo on complete single-study datasets, secondary analyses of published clinical trials present a distinctly more complex challenge. Critical biomarker stratifications are frequently reported across multiple publications over time, leading to asynchronous follow-up durations. Furthermore, biomarker data are often intrinsically fragmented. For example, while programmed death-ligand 1 (PD-L1) expression is typically assessed in the full intent-to-treat (ITT) population, tumor mutational burden (TMB) evaluation is strictly confined to a subset of patients with available

tissue. Such spatial (population subsets) and temporal (follow-up) misalignments pose significant challenges to conventional summary-based reconstruction methods. To demonstrate how MD-JoPiGo harmonizes cross-publication data and resolves hidden intersectional topologies, we validated the framework using asynchronous reports from the CheckMate 227 trial.

We aggregated 1D marginal event-free survival (EFS) constraints from two independent publications. The overall EFS and evaluable TMB-stratified trajectories were reconstructed from published Kaplan-Meier (KM) images in an earlier report[20]. The full-population PD-L1 stratification curves (both $\geq 1\%$ and $< 1\%$) were obtained from a subsequent update with a longer follow-up period[21]. To resolve temporal discrepancies, we applied administrative right-censoring to truncate the extended PD-L1 trajectories at 24 months, ensuring aligned time horizons across all inputs. Although the earlier report [Insert Citation 1] provided the actual EFS outcomes for the joint biomarker subpopulations (e.g., TMB-High concurrent with PD-L1 $\geq$ 1% or $< 1\%$), these intersectional trajectories were excluded from the reconstruction process. MD-JoPiGo synthesized the multidimensional clinical profiles using only the 1D margins, and the published joint outcomes served as a blinded ground-truth validation set.

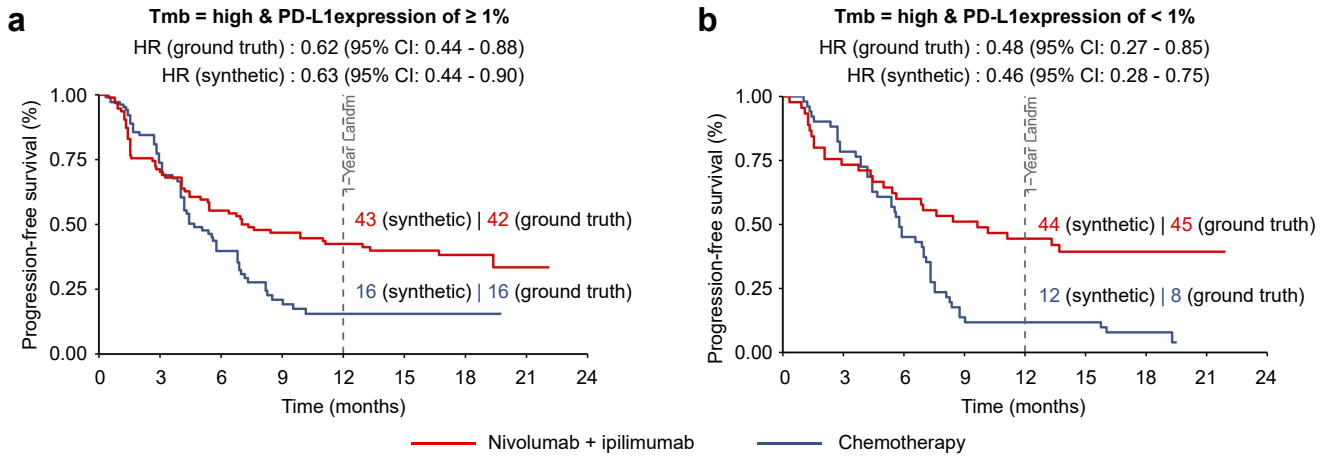

Fig. 5 | Validation of multi-arm cross-survival trajectories synthesized by MD-JoPiGo in multi-biomarker intersecting subgroups. a, Survival comparison between Nivolumab + Ipilimumab and Chemotherapy within the TMB-High and PD-L1 $\geq$ 1% subpopulation. The synthetic digital twin cohort demonstrated high concordance with the literature-derived ground truth. The ground truth hazard ratio (HR) is 0.62 (95% CI: 0.44–0.88), compared to the synthetic HR of 0.63 (95% CI: 0.44–0.90). The landmark 1-year survival rates for the synthetic cohort were 43% (Nivolumab + Ipilimumab) and 16% (Chemotherapy), strictly mirroring the reported clinical ground truth of 42% and 16%, respectively. **b**, Survival comparison within the implicitly reconstructed TMB-High and PD-L1 $< 1\%$ subpopulation. Despite the absence of explicit marginal survival constraints for the PD-L1 $< 1\%$ group, MD-JoPiGo accurately recovered the intersectional survival dynamics. The synthetic HR is 0.46 (95% CI: 0.28–0.75), aligning closely with the ground truth HR of 0.48 (95% CI: 0.27–0.85). The synthetic 1-year survival rates were 44% (Nivolumab + Ipilimumab) and 12% (Chemotherapy), consistent with the reported literature benchmark of 45% and 8%.

Despite relying on an incomplete TMB subset and time-truncated PD-L1 populations, MD-JoPiGo reconstructed the multidimensional event-free survival (EFS) trajectories for both intersectional subgroups. Within the TMB-High and PD-L1 $\geq$ 1% subpopulation, the synthetic Nivolumab plus Ipilimumab versus Chemotherapy comparison yielded a hazard ratio (HR) of 0.63 (95% CI: 0.44–0.90), consistent with the literature-derived ground truth of 0.62 (95% CI: 0.44–0.88). The overall trajectories of the synthesized curves were visually consistent with the original published Kaplan-Meier plots[20]. Furthermore, the 1-year EFS rates for the synthetic cohort (43% vs. 16%) aligned with the clinical benchmarks (42% vs. 16%). For the TMB-High and PD-L1 $< 1\%$ subpopulation, the framework yielded a synthetic HR of 0.46 (95% CI: 0.28–0.75), comparable to the clinical benchmark of 0.48 (95% CI: 0.27–0.85), with synthesized curve trends also reflecting the published data. The reconstructed 1-year EFS rates in this subgroup (44% vs. 12%) approximated the reported literature values (45% vs. 8%).

To evaluate the framework under extreme data sparsity, we conducted a stress test by withholding the explicit PD-L1 $<$ 1% marginal curves, providing the algorithm solely with the ITT and PD-L1 $\geq 1\%$ margins (Supplementary Figure 6). In this scenario, MD-JoPiGo estimated the intersectional dynamics through residual topological space. The synthetic HR for the TMB-High and PD-L1 $\geq 1\%$ cohort was 0.59 (95% CI: 0.41–0.85), with 1-year EFS rates of 46% and 20%. Within the unconstrained TMB-High and PD-L1 $< 1\%$ subgroup, the implicitly deduced HR was 0.48 (95% CI: 0.29–0.78), matching the clinical ground truth of 0.48. While the absolute 1-year EFS probabilities showed wider variances (36% vs. 4%) due to the relaxed constraint boundaries, the relative treatment hazard topology was preserved.

Discussion

MD-JoPiGo synthesizes multidimensional clinical profiles from standard survival summaries. One-dimensional individual patient data (IPD) are first extracted from Kaplan–Meier plots using the KM-PoPiGo tool. The framework then models these marginals into a high-dimensional joint distribution[22]. In our evaluations where prognostic factors were relatively independent, the synthetic profiles matched the reference data. For cases involving overlapping factors—such as age and physical frailty—incorporating minimal structural priors maintained statistical stability and prevented risk misallocation. This approach facilitates the estimation of intersectional treatment effects not explicitly reported in the source literature[23].

The coefficient drift observed in the lung cancer analysis reflects the challenge of structural identifiability, which is determined by the underlying causal topology of the clinical covariates[19]. When prognostic factors operate as parallel predictors—such as age, sex, and node status in the colorectal cancer dataset—conditional independence given survival time is maintained. The framework leverages these complementary signals to narrow the feasible solution space. Conversely, when covariates form a confounding mediation structure, such as age driving physical frailty, the downstream mediator blocks the prognostic information of the upstream variable. Under such severe collinearity, the Fisher information matrix approaches singularity, rendering the joint distribution structurally unidentifiable from marginal survival data alone[24]. Similarly, when trial selection criteria act as a collider, conditioning on enrollment induces spurious negative correlations between independent predictors (Berkson's paradox)[25]. This distortion of the joint space further necessitates structural priors.

This causal ambiguity explains why the independence assumption inherent in standard maximum entropy (MaxEnt) models succeeds or fails depending on the underlying topology. Consequently, pre-screening covariate structures using Directed Acyclic Graphs (DAGs)[19] is a necessary step prior to synthesis. While parallel predictors are recovered under default MaxEnt assumptions, complex topologies require structural priors. For chain mediations, synthesizing only the terminal variable or anchoring the synthesis with minimal priors (e.g., joint proportions) resolves localized coefficient drift. Similarly, mitigating the spurious negative correlations induced by collider selection requires these same structural anchors. From a clinical perspective, estimating intersectional treatment effects supports the emulation of complex trial scenarios where published data are limited. In oncology, evaluating efficacy across individual-level subcohorts informs personalized treatment decisions[26]; however, RCT publications rarely disclose these effects due to space constraints and sample size attrition within highly stratified subsets[27]. In practice, prognostic constraints such as biomarker stratifications are often fragmented across multiple, temporally misaligned publications. The CheckMate 227 analysis demonstrates that for biomarkers operating as parallel predictors, integrating these asynchronous reports allows for the robust estimation of joint survival outcomes, enabling secondary analyses that span multiple literature sources. MD-JoPiGo estimates these multivariable relationships from standard marginal reports, supporting the construction of Synthetic Control Arms (SCA)[28] and external control validations. This alignment of real-world patient profiles with trial-specific subcohorts extends the utility of published clinical evidence.

Reliable trial emulation[29] from marginal summaries depends on more transparent reporting standards. The results demonstrate that providing a single intersectional summary—such as the joint proportion of elderly patients with impaired performance status—is sufficient to correct the bias from entropy-driven risk redistribution. We therefore recommend that future RCT reports include joint counts for clinically coupled baseline characteristics, as encouraged by the CONSORT 2010 guidelines[30]. This minimal structural information preserves patient privacy while supporting the growing demand for collaborative privacy-preserving oncological analytics[31] and precise synthetic trial emulation.

Synthesis fidelity is constrained by the information density of the input marginal summaries. Although the CheckMate 227 stress test demonstrated that relative treatment efficacy can be robustly preserved even under extreme data sparsity (e.g., withholding explicit marginal constraints), accurately recovering absolute survival probabilities remains highly sensitive to the completeness of the input marginal boundaries. Future research could leverage Large Language Models (LLMs)[32] to extract unstructured constraints from trial full-texts or integrate empirical priors from large-scale registries. However, mitigating the risk of numerical hallucination necessitates a semi-automated framework where AI-driven extraction is paired with expert validation to ensure that supplemental priors remain statistically sound and clinically valid.

MD-JoPiGo synthesizes multidimensional clinical profiles by integrating 1D-IPD data reconstructed from published KM curve images across both overall and stratified populations. By accounting for the structural dependencies between clinical factors, the framework recovers the individual-level detail required for intersectional treatment analysis without requiring access to original trial datasets. This methodology establishes a scalable foundation for advanced evidence synthesis, including multicenter data integration, individual patient data (IPD) meta-analysis, and the development of multivariable prognostic models.

Methods

Empirical datasets and clinical scenarios

To evaluate the structural identifiability and synthesis fidelity of the MD-JoPiGo framework across diverse causal topologies, we analyzed independent, publicly available oncology datasets. These datasets provide the complete multivariable joint distributions required to establish a ground truth reference for intersectional survival dynamics.

Colon cancer cohort (Parallel independence) We analyzed data from a Phase III randomized controlled trial evaluating fluorouracil plus levamisole in stage III colon cancer ($N = 929$)[18]. This cohort provides a baseline for the *parallel independence* causal topology (Scenario A). In this dataset, baseline prognostic factors—such as sex, histological grade, and nodal involvement—act as independent predictors of disease-free survival. Because these variables lack severe structural collinearity, we can test the framework's ability to resolve multivariable dynamics using default maximum entropy assumptions, without supplemental structural priors.

Lung cancer cohort (Chain mediation) To evaluate the framework under interdependent structural constraints, we analyzed an advanced lung cancer dataset from the North Central Cancer Treatment Group ($n = 228$)[17]. This cohort represents a *chain mediation* topology (Scenario B). Clinically, advanced age contributes to physical frailty, such that age influences overall survival primarily through the mediating variable of ECOG performance status. This structural dependency typically induces localized coefficient drift during unconstrained synthesis. We used this dataset to demonstrate that incorporating a minimal structural prior (e.g., a single joint proportion of age and performance status) corrects this bias and recovers the reference multivariable profiles.

CheckMate 227 trial cohort (Fragmented and asynchronous data) We analyzed asynchronous publications from the CheckMate 227 trial to evaluate the framework's resolution of hidden intersectional topologies from fragmented real-world reporting. These asynchronous reports present spatial misalignments: programmed death-ligand 1 (PD-L1) expression was assessed in the full intent-to-treat population[21], whereas tumor mutational burden (TMB) evaluation was confined to a patient subset[20]. Varying follow-up durations across the publications additionally introduced temporal discrepancies. To align the time horizons of the aggregated one-dimensional marginal event-free survival (EFS) constraints, we applied administrative right-censoring at 24 months. The published EFS outcomes for the joint biomarker subpopulations were explicitly excluded from the reconstruction process and reserved as a blinded ground-truth validation set.

For these empirical cohorts, the reference data were used to generate KM images for the overall population and 1D-stratified subgroups. These images were then digitized using the open-source KM-PoPiGo web tool (`https://kmpopigo.github.io`) to reconstruct one-dimensional individual patient data (1D-IPD) as standard CSV files. This process simulates the real-world workflow of extracting and synthesizing multidimensional clinical profiles from the dimensionally reduced survival summaries typical of published study.

Extraction of one-dimensional individual patient data

The extraction of 1D-IPD is formulated as a constraint-consistent inverse optimization problem. Serving as the preprocessing module, KM-PoPiGo employs joint global optimization to estimate individual time-to-event and censoring coordinates. The algorithm adapts to the data granularity of the source study by integrating up to four specific reporting constraints: total sample size, number-at-risk tables, total event counts, and explicit censoring times. The resulting 1D-IPD sets, comprising (t_i, δ_i) tuples per patient, provide the marginal inputs for the subsequent multidimensional reconstruction.

Multidimensional joint distribution reconstruction

Problem Formulation Let D_{total} represent the reconstructed IPD derived from the overall KM curve. Each patient $i \in \{1, \ldots, N\}$ is characterized by an observed tuple (t_i, δ_i), representing realizations of the time-to-event (T) and event indicator (Δ) random variables. The objective is to determine a mapping $F : i \to \boldsymbol{x}_i$, where $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,D}]^\top$ is a realization of the multidimensional feature vector $\mathbf{X}$ encompassing D clinical characteristics (e.g., disease stage, histology, and biomarker expression). We seek to reconstruct an empirical patient-level state matrix $\boldsymbol{\Theta}$ that preserves the underlying joint distribution $P(\mathbf{X}, T, \Delta)$ and satisfies the marginal survival constraints reported in the source study.

Joint distribution estimation via maximum entropy Prior to assigning individual clinical profiles, the joint probability distribution of the cohort must be estimated. Let $\mathscr{X}$ denote the state space comprising all possible combinations of the D clinical characteristics. For any multidimensional state $\boldsymbol{x} \in \mathscr{X}$, the objective is to estimate its joint probability $P(\boldsymbol{x})$.

According to Jaynes' principle of maximum entropy[33], the most unbiased representation of the joint distribution maximizes Shannon entropy, subject to known empirical constraints. This is formulated as a constrained optimization problem:

$$\max_P H(P) = -\sum_{\boldsymbol{x}\in\mathscr{X}} P(\boldsymbol{x}) \ln P(\boldsymbol{x}) \tag{1}$$

subject to the probability axioms:

$$\sum_{\boldsymbol{x}\in\mathscr{X}} P(\boldsymbol{x}) = 1, \quad P(\boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \mathscr{X} \tag{2}$$

and the reconstructed 1D marginal constraints:

$$\sum_{\boldsymbol{x}\in\mathscr{X}:x_d=v} P(\boldsymbol{x}) = P(X_d = v) \quad \forall v \in \mathscr{V}_d, \text{ for } d = 1,2,\ldots,D \tag{3}$$

where X_d represents the d-th clinical variable, $\mathscr{V}_d$ defines its corresponding set of possible states (e.g., $\{\text{male}, \text{female}\}$), and $P(X_d = v)$ denotes the observed marginal probability for that specific state.

The framework adapts to the underlying causal topology by incorporating structural priors. Under *parallel independence*, the optimization relies solely on the 1D marginal constraints (Eq. 3), which naturally yields conditional independence under the maximum entropy objective. To resolve structural unidentifiability in *chain mediation* or *collider* topologies, the optimization requires supplemental structural constraints, such as a known joint probability π_{ab} between specific coupled variables X_a and X_b:

$$\sum_{\boldsymbol{x}\in\mathscr{X}:x_a=v_a,x_b=v_b} P(\boldsymbol{x}) = \pi_{ab}(v_a, v_b) \quad \text{for specific } (v_a, v_b) \in \mathscr{V}_a \times \mathscr{V}_b \tag{4}$$

Solving this constrained optimization yields the optimal joint probability distribution $P^*(\boldsymbol{x})$. Scaling these probabilities by the total cohort size N determines the target patient count, $N_{\boldsymbol{x}} \approx N \cdot P^*(\boldsymbol{x})$, for each multidimensional state $\boldsymbol{x} \in \mathscr{X}$. These discrete allocations establish the definitive combinatorial constraints for the subsequent individual-level label exchange process.

Objective function formulation The individual-level reconstruction is formulated as a combinatorial optimization problem. We define the objective function $J(\boldsymbol{\Theta})$ to quantify the discrepancy between the synthesized and reference marginal survival curves using the integrated squared error (ISE):

$$J(\boldsymbol{\Theta}) = \sum_{g\in\mathscr{G}} \int_0^{T_{\max}} \left(\hat{S}_g(t \mid \boldsymbol{\Theta}) - S_g^{\text{target}}(t)\right)^2 dt \tag{5}$$

where:

- $\boldsymbol{\Theta}$ is the patient-level state matrix of clinical feature assignments.
- $\mathscr{G}$ represents the set of all available 1D marginal subgroups (e.g., the overall cohort, male, female, age ≥ 65) serving as optimization targets.
- $S_g^{\text{target}}(t)$ is the reference survival probability for subgroup $g \in \mathscr{G}$, derived from the initial 1D-IPD extraction.
- $\hat{S}_g(t \mid \boldsymbol{\Theta})$ is the Kaplan-Meier survival estimate computed from the synthesized cohort subset corresponding to subgroup g under the state matrix $\boldsymbol{\Theta}$.

Optimization via simulated annealing To minimize the non-convex objective function $J(\boldsymbol{\Theta})$, simulated annealing (SA) is employed. The search space is explored by randomly selecting two patients and exchanging their multidimensional clinical feature vectors. This pairwise exchange mechanism ensures that both the predefined joint distribution constraints ($N_{\boldsymbol{x}}$) and the overall marginal frequencies are strictly preserved throughout the optimization.

The acceptance probability for a candidate state transition $\boldsymbol{\Theta} \to \boldsymbol{\Theta}'$ is governed by the Metropolis criterion:

$$P_{\text{acc}} = \min\left(1, \exp\left(-\frac{J(\boldsymbol{\Theta}') - J(\boldsymbol{\Theta})}{T_k}\right)\right) \tag{6}$$

where T_k denotes the system temperature at iteration k, dynamically reduced according to a geometric cooling schedule $T_{k+1} = \eta T_k$. Following empirical calibration, the algorithm is initialized at $T_0 = 2.0$ with a cooling rate of $\eta = 0.9998$. The optimization executes for $50,000$ iterations, ultimately returning the optimal state matrix $\boldsymbol{\Theta}^*$ that minimizes the objective function.

Multivariable statistical inference Upon convergence, the synthesized multidimensional clinical profiles (D_{syn}) are evaluated using multivariable Cox proportional hazards regression to estimate adjusted hazard ratios (aHRs):

$$h(t \mid Z, \boldsymbol{x}) = h_0(t) \exp\left(\beta_{\text{treat}} Z + \sum_{d=1}^{D} \beta_d x_d\right) \tag{7}$$

where $h_0(t)$ is the baseline hazard, Z denotes the treatment indicator, and x_d represents the d-th component of the reconstructed baseline covariate vector $\boldsymbol{x}$ derived from the optimal state matrix $\boldsymbol{\Theta}^*$.

Simulation framework and causal topologies

To evaluate the reconstruction fidelity of MD-JoPiGo, we conducted Monte Carlo simulations ($n = 200$ iterations per scenario) to generate synthetic cohorts comprising $N = 200$ individuals each. For every simulated patient, two binary predictors (X_1, X_2) and a survival outcome (time-to-event T and censoring indicator Δ) were generated. Survival times were drawn from a Cox-Weibull model:

$$h(t \mid X_1, X_2) = \lambda \gamma t^{\gamma - 1} \exp(\beta_1 X_1 + \beta_2 X_2) \tag{8}$$

with baseline parameters set at scale $\lambda = 0.01$ and shape $\gamma = 1.2$. The target hazard ratios (HRs) were defined as $\text{HR}_{X_1} = 0.5$ ($\beta_1 = \ln 0.5 \approx -0.693$) and $\text{HR}_{X_2} = 0.75$ ($\beta_2 = \ln 0.75 \approx -0.288$). Independent right-censoring times were generated from an exponential distribution, calibrated to maintain an overall censoring rate of approximately 20%. The algorithm's performance was benchmarked across three fundamental causal topologies:

- **Parallel independence (scenario A):** X_1 and X_2 were modeled as independent and identically distributed (i.i.d.) Bernoulli trials ($p = 0.5$). In this baseline scenario, we evaluated the default MaxEnt mode (relying solely on 1D marginals) and compared its performance against a structurally calibrated reference (utilizing true joint frequencies). This comparison demonstrates that for independent predictors, the MaxEnt assumption of conditional independence inherently matches the data structure, obviating the need for supplemental structural priors.
- **Chain mediation (scenario B):** A directed dependency was introduced ($X_1 \to X_2$), where $X_1 = 1$ increased the probability of $X_2 = 1$, inducing a Pearson's correlation $\phi \approx 0.5$. To perform structural calibration, MD-JoPiGo was provided with the 2×2 contingency table of X_1 and X_2. This prior information enables the algorithm to bypass the default independence assumption and allocate patient profiles according to the true joint probability.
- **Collider selection (scenario C):** X_1 and X_2 remained independent in the source population, but inclusion in the cohort was conditioned on a collider C ($X_1 \to C \leftarrow X_2$). Only individuals with $C = 1$ were retained in the analyzed sample, inducing an artificial negative correlation (Berkson's paradox)[25]. Because the pre-selection ground truth is inherently unobservable in empirical randomized controlled trial (RCT) datasets—which already represent a collider-conditioned population—we relied strictly on this simulation to evaluate synthesis fidelity under selection bias. Structural calibration was implemented by inputting the joint distribution observed within the selected sample. This ensures the reconstructed IPD preserves the trial's internal multivariable dynamics, preventing secondary synthesis distortion.

For each topology, we compared the multivariable HRs derived from the ground truth sample against those estimated from the default MaxEnt and structurally calibrated reconstructions.

Statistical analysis and evaluation metrics

All statistical analyses and IPD reconstructions were performed using the R statistical computing environment (version 4.3.3; R Foundation for Statistical Computing). Survival distributions were estimated using the Kaplan-Meier method, and multivariable prognostic dynamics were analyzed using the Cox proportional hazards regression model (via the `survival` package[34]).

To quantify the performance of the MD-JoPiGo framework, we evaluated the synthetic cohorts against the original ground-truth data across two primary dimensions: marginal survival fidelity and multivariable intersectional consistency.

Statistical validation of survival concordance The synthesis fidelity of the MD-JoPiGo framework was assessed via mathematical divergence metrics and statistical hypothesis testing. First, the deviation between the synthesized and reference marginal survival curves was quantified using the integrated squared error (ISE) for each subgroup $g \in \mathscr{G}$:

$$\text{ISE}_g = \int_0^{T_{\max}} \left(\hat{S}_g(t \mid \boldsymbol{\Theta}) - S_g^{\text{target}}(t)\right)^2 dt \tag{9}$$

Second, to evaluate the statistical concordance between the synthesized multidimensional clinical profiles and the ground truth data, two-sided log-rank tests were employed across all stratified subpopulations. Rather than proving absolute equivalence, a resulting P-value > 0.05 was interpreted as an absence of statistically significant divergence, serving as a supplementary heuristic to the primary ISE metric to confirm synthesis fidelity.

Multivariable intersectional consistency To assess the framework's capacity to recover unobserved high-order interactions, the adjusted hazard ratios (aHRs) derived from the synthesized multidimensional clinical profiles were benchmarked against those obtained from the ground truth patient profiles. The relative error (RE) for each estimated log-hazard coefficient was quantified as:

$$\mathrm{RE}(\hat{\beta}_d) = \left| \frac{\hat{\beta}_d^{\mathrm{syn}} - \beta_d^{\mathrm{true}}}{\beta_d^{\mathrm{true}}} \right| \times 100\% \tag{10}$$

where $\hat{\beta}_d^{\mathrm{syn}}$ represents the coefficient estimated from the synthesized profiles, and β_d^{true} is the reference parameter. Furthermore, structural identifiability across varying causal topologies was evaluated by assessing whether the 95% confidence intervals (CIs) computed from the synthesized profiles successfully encompassed the ground truth multivariable aHRs.

Code availability

The complete source code for the MD-JoPiGo framework, including the maximum entropy estimation and simulated annealing optimization modules, is open-source and publicly available on GitHub at https://github.com/ZheqingZhu/mdjopigo.

Declarations

Ethical approval

This study constitutes a secondary analysis of de-identified, publicly available data and summary statistics extracted from published clinical trials. No new human participants were recruited, and no clinical interventions were performed. Consequently, Institutional Review Board (IRB) approval and patient informed consent were deemed not applicable and were waived. All data synthesis and reconstruction procedures were conducted in accordance with the ethical principles outlined in the Declaration of Helsinki.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information is available for this paper.

References

1. Kent, D. M. & Hayward, R. A. Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *JAMA* **298**, 1209–1212 (2007).
2. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *New Engl. J. Medicine* **372**, 793–795 (2015).
3. Selby, J. V., Maas, C. C. H. M., Fireman, B. H. & Kent, D. M. Predictive modeling of heterogeneous treatment effects in RCTs: A scoping review. *JAMA Netw. Open* **8**, e2522390 (2025).
4. Blay, J.-Y. *et al.* Pembrolizumab in patients with rare and ultra-rare sarcomas (AcSé Pembrolizumab): analysis of a subgroup from a non-randomised, open-label, phase 2, basket trial. *The Lancet Oncol.* **24**, 892–902 (2023).
5. Wang, E. H. *et al.* Crossing borders securely: synthetic data and federated networks for privacy-preserving access to real-world data and emerging use cases. *npj Digit. Medicine* **8**, 758 (2025).
6. Liu, S.-Y. M. *et al.* First-line pyrotinib in advanced HER2-mutant non-small-cell lung cancer: a patient-centric phase 2 trial. *Nat. Medicine* **29**, 2079–2086 (2023).
7. Jung, R. G. *et al.* Prognostic factors associated with mortality in septic shock: a systematic review and meta-analysis. *The Lancet Respir. Medicine* **14**, 225–232 (2026).

8. Powles, T. *et al.* Enfortumab vedotin and pembrolizumab in untreated advanced urothelial cancer. *New Engl. J. Medicine* **390**, 875–888 (2024).
9. Yang, J. C.-H. *et al.* Overall survival with amivantamab–lazertinib in EGFR-mutated advanced NSCLC. *New Engl. J. Medicine* **393**, 1681–1693 (2025).
10. Lee, S.-H. *et al.* Lazertinib versus osimertinib in previously untreated EGFR-mutant advanced NSCLC: A randomized, double-blind, exploratory analysis from MARIPOSA. *J. Thorac. Oncol.* **20**, 1655–1668 (2025).
11. Hellmann, M. D. *et al.* Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *New Engl. J. Medicine* **378**, 2093–2104 (2018).
12. Smit, J. M. *et al.* Predicting benefit from adjuvant therapy with corticosteroids in community-acquired pneumonia: a data-driven analysis of randomised trials. *The Lancet Respir. Medicine* **13**, 221–233 (2025).
13. Fisher, D. J., Carpenter, J. R., Morris, T. P., Freeman, S. C. & Tierney, J. F. Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* **356**, j573 (2017).
14. Guyot, P., Ades, A. E., Ouwens, M. J. N. M. & Welton, N. J. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* **12**, 9 (2012).
15. Liu, N., Zhou, Y. & Lee, J. J. IPDfromKM: reconstruct individual patient data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* **21**, 111 (2021).
16. Song, Y., Sun, F., Redline, S. & Wang, R. Random-effects meta-analysis of combined outcomes based on reconstructions of individual patient data. *Res. Synth. Methods* **11**, 594–616 (2020).
17. Loprinzi, C. L. *et al.* Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *J. Clin. Oncol.* **12**, 601–607 (1994).
18. Moertel, C. G. *et al.* Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New Engl. J. Medicine* **322**, 352–358 (1990).
19. Greenland, S., Pearl, J. & Robins, J. M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
20. Hellmann, M. D. *et al.* Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *New Engl. J. Medicine* **378**, 2093–2104 (2018).
21. Paz-Ares, L. G. *et al.* First-line nivolumab plus ipilimumab in advanced nsclc: 4-year outcomes from the randomized, open-label, phase 3 checkmate 227 part 1 trial. *J. Thorac. Oncol.* **17**, 289–308 (2022).
22. Fezzi, S. *et al.* Individual patient data meta-analysis of paclitaxel-coated balloons vs. drug-eluting stents for small-vessel coronary artery disease: the andromeda study. *Eur. Hear. J.* **46**, 1586–1599 (2025).
23. Tan, D. J. H. *et al.* Survival trends in sorafenib for advanced hepatocellular carcinoma: A reconstructed individual patient data meta-analysis of randomized trials. *Liver Cancer* **12**, 445–456 (2023).
24. Rothenberg, T. J. Identification in parametric models. *Econom. J. Econom. Soc.* 577–591 (1971).
25. Hernán, M. A., Hernández-Díaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
26. Gerritsen, J. K. W. *et al.* Practical and statistical aspects of subgroup analyses in surgical neuro-oncology: A comprehensive review from the pioneer consortium. *Neuro-Oncology* **27**, 1149–1164 (2024).
27. Sherry, A. D. *et al.* Differential treatment effects of subgroup analyses in Phase 3 oncology trials from 2004 to 2020. *JAMA Netw. Open* **7**, e243379 (2024).
28. Serrano, C. *et al.* Rethinking placebos: embracing synthetic control arms in clinical trials for rare tumors. *Nat. Medicine* **29**, 2689–2692 (2023).
29. Hernán, M. A. & Robins, J. M. Using big data to emulate a Target Trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
30. Schulz, K. F., Altman, D. G., Moher, D. & CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c332 (2010).
31. Geva, R., Gusev, A., Polyakov, Y. *et al.* Collaborative privacy-preserving analysis of oncological data using multiparty homomorphic encryption. *Proc. Natl. Acad. Sci.* **120**, e2304415120 (2023).
32. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Medicine* **29**, 1930–1940 (2023).

33. Jaynes, E. T. Information theory and statistical mechanics. *Phys. review* **106**, 620 (1957).

34. Therneau, T. M., Lumley, T., Atkinson, E. & Crowson, C. *survival: Survival Analysis* (2024). R package version 3.8-3.