Chilean Journal of Statistics
Vol. 16, No. 2, December 2025, 125–177
DOI: 10.32372/ChJS.16-02-02

Survival Analysis
Research Article

Non-Markov multi-state survival analysis with complex censoring: A structured synthesis of models, estimators, and applications

Marta Azevedo[1,*], Luís Meira-Machado[1], and Carla Moreira[1]
[1]Centre of Mathematics, Universidade do Minho, Braga, Portugal

(Received: 30 July 2025 · Accepted: 01 September 2025)

Abstract

Reliable quantification of treatment benefit in late-phase clinical trials increasingly requires modeling patient histories that include progression, adverse events, and treatment switches. Conventional multi-state analyses often invoke the Markov property and assume independent right censoring— conditions rarely satisfied in oncology, immunology, or cell-therapy programs, where intermediate events and informative dropout are common. This article presents a systematic review and bibliometric synthesis of 48 peer-reviewed studies published through 11 June 2025 that (i) relax the Markov assumption and (ii) address complex observation schemes such as left truncation, interval censoring, or informative censoring, identified through Web of Science and Scopus searches following preferred reporting items for systematic reviews and meta-analyses 2020 guidelines.A recurring set of methodological strategies emerges across the literature, including semi-Markov transition-intensity models, illness–death and semi-competing risks frameworks, landmarking for dynamic prediction, and inverse-probability-of-censoring weighting. Estimation approaches range from nonparametric product integrals to semiparametric weighted likelihoods and Bayesian Markov chain Monte Carlo, with recent contributions exploring saddle-point approximations and subsampling for large-scale electronic health records. To complement this synthesis, we include a compact simulation contrasting baseline and landmark Aalen–Johansen estimators under semi-Markov dynamics with history-dependent censoring, and a bibliometric network analysis mapping collaboration patterns, thematic clusters, and structural gaps. The findings highlight the need for scalable, auditable software, robust diagnostics aligned with the International Council for Harmonization E9(R1) estimand framework (which links clinical trial objectives to precise statistical targets), and better integration of high-dimensional biomarkers; limitations include the English-language restriction and reliance on bibliometric metadata. Addressing these priorities may enhance both the methodological robustness and regulatory applicability of non-Markov survival models.

Keywords: History-dependent censoring · Interval/panel observation · Left truncation · Non-Markov inference · PRISMA · Pseudo-observations

Mathematics Subject Classification: Primary 62N02 · Secondary 62P10

*Corresponding author. Email: marta.vasconcelos4@gmail.com (M. Azevedo)

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

1. Introduction

The quantitative analysis of time-to-event outcomes has long supported clinical decision-making, benefit–risk assessment, and the regulatory approval of new therapies [1, 2]. Traditional survival models typically assume a single terminal event. When extended to multi-state settings, they often adopt the Markov property, whereby future hazards depend only on the current state and not on the history leading to it [3]. However, in modern applications—such as oncology trials tracking progression followed by death, transplantation studies involving graft loss prior to mortality, or real-world analyses of electronic health records (EHRs) with recurrent hospitalizations—patients often experience sequences of events whose timing alters future risks. Ignoring this history can bias treatment effect estimates, reduce statistical power, and complicate estimand definition under the International Council for Harmonization E9(R1) estimand framework [4, 5].

Multi-state survival models provide a natural extension of the Cox framework by allowing transitions among multiple states, offering a more granular representation of disease dynamics [6, 7]. Nonetheless, many applied analyses continue to rely on a (time-inhomogeneous) Markov assumption and to restrict attention to independent (non-informative) right censoring. In contrast, real-world datasets frequently exhibit more complex observation structures: delayed entry induces left truncation; irregular follow-up leads to interval censoring; and non-random dropout—often due to adverse events—can render censoring informative [8, 9]. These features challenge the assumptions of Markov models and motivate the use of non-Markov approaches capable of incorporating history dependence, semi-competing risks, and dynamic prediction [10].

In pivotal oncology studies, the illness–death structure linking progression-free survival (PFS) and overall survival (OS) is often more appropriately modeled using a semi-Markov framework, as the post-progression hazard typically depends on the time already spent in the progressed state. Both simulation studies and empirical analyses have shown that, under path dependence, standard Aalen–Johansen (AJ) estimators can produce biased conditional transition probabilities defined after a landmark time, and they are also biased when censoring is history-dependent (informative) [11, 12, 13]. By contrast, for unconditional transition probabilities measured from study entry and under non-informative censoring, AJ remains a valid baseline estimator. In this case, it consistently recovers the transition matrix of the Markovized process obtained by averaging the history-dependent transition intensities among individuals currently in each state, even when the underlying process is non-Markov [3, 12, 14, 15]. Similar methodological challenges arise in adaptive trials with treatment switching or crossover designs, particularly in rare disease contexts, where censoring mechanisms are frequently informative by design.

Recent contributions have addressed these limitations through a range of methodological innovations, including semi-Markov transition-intensity models, illness–death and semi-competing risks structures (nonterminal events may be censored by a terminal event, but not vice-versa), landmarking for time-updated prediction, and inverse-probability-of-censoring weighting (IPCW) —a special case of inverse probability weighting (IPW)— to mitigate informative censoring. Estimation strategies span nonparametric product integrals, semiparametric weighted likelihoods, and Bayesian Markov chain Monte Carlo (MCMC), with some studies introducing computational refinements —such as saddle-point approximations, parallel computing, and rare-event subsampling— to improve scalability in large-scale EHR–based cohorts [16]. However, the resulting literature is fragmented across journals, methodological frameworks, and software platforms, making it difficult for practitioners and applied researchers to form a coherent picture of the field. In particular, no prior work to our knowledge offers an integrated synthesis that combines technical review of non-Markov estimation methods with a bibliometric mapping of research activity and thematic structure.

In this article, we address this gap through two integrated contributions. First, we provide a structured synthesis of models and estimators for non-Markov multi-state survival under complex observation schemes (left truncation, interval/irregular censoring, and informative dropout), linking theory to practice via product-integral representations, influence-function–based variance, and IPCW adjustments.

Second, we construct a quantitative bibliometric map of the field —covering 48 peer-reviewed studies screened to 11 June 2025 under preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020— to characterize collaboration patterns, thematic clusters, and structural gaps [17]. What is new relative to prior tutorials and narrative reviews (such as Markov-centric overviews) is that we (i) focus explicitly on non-Markov dynamics and censoring complexity in a single, unified notation; (ii) translate the synthesis into practitioner-facing decision rules (Table 3), diagnostics, and a curated software landscape (Table 12); (iii) align recommendations with the ICH E9(R1) estimand framework to support regulatory use; and (iv) provide reproducible materials (Appendices A, B) and a compact simulation isolating semi-Markov effects and history-dependent censoring that clarifies where baseline versus landmark AJ (LAJ) will —and will not— perform well. These strands highlight both methodological advances and the remaining obstacles to wider adoption —most notably the need for scalable, auditable software and routine diagnostics— while offering concrete guidance for applied teams.

This review addresses the following research questions (RQs), which guide the methodological synthesis, simulation, and bibliometric analysis:

RQ1 Methods and estimands —Which estimators and modeling frameworks consistently relax the Markov assumption and handle complex observation schemes (left truncation, interval/irregular censoring, informative dropout), and under what conditions do they apply? (See Section 2; Table 3.)

RQ2 Finite-sample performance —In a canonical progressive illness–death setting with semi-Markov post-progression hazards and history-dependent censoring, how do baseline AJ and LAJ compare in terms of bias and root mean square error (RMSE) for conditional transition probabilities between a landmark and a future time? (See Section 2.4; Table 4.)

RQ3 Software readiness —What open-source `R` tooling currently supports these analyses, and what assumptions, strengths, and limitations matter in practice? (See Table 12.)

RQ4 Field structure and gaps —What collaboration patterns and thematic clusters characterize this literature, and where are the structural gaps? (See Section 4; Figures 4 and 7.)

RQ5 Implementation priorities — Which diagnostics and workflow standards (including alignment with the ICH E9(R1) estimand framework) should be prioritized to support robust, auditable analyses at scale? (See Section 5.)

To clarify how this review differs from prior surveys, Table 1 contrasts the scope and coverage of widely cited reviews with the present article. Relative to earlier work, our contribution integrates non-Markov estimation under left truncation, interval/irregular observation and informative censoring, links these methods to IPCW and diagnostics aligned with ICH E9(R1) estimands, and adds a software map, a bibliometric synthesis, and a focused simulation to illustrate finite-sample behavior.

Table 1.: Comparison of the present review with prior surveys on multi-state modeling.

Review	Non-Markov	LT / IC/InfC	Transition probability	Land-marking	IPCW	Pseudo-obs./GEE	Software map	Biblio-metrics	Notes on scope/emphasis
Putter et al. (2007) [6]	△	△/△/✗	△	△	✗	✗	△	✗	Tutorial on multi-state (Markov baseline), product integrals, data prep; limited non-Markov and censoring complexity.
Jackson (2011) [7]	△	△/△/✗	✗	△	✗	✗	✓	✗	`msm` package paper (panel data, Markov focus); implementation and prediction for Markov settings.
Andersen et al. (2012) [3]	△	△/△/✗	△	△	✗	△	✗	✗	Competing risks/multi-state pitfalls; some non-Markov discussion; limited treatment of informative censoring.
Putter and Spi-toni (2018)	✓	△/△/✗	✓	✓	△	△	✗	✗	Focus on non-Markov transition probabilities (landmark versus plug-in); less on complex censoring and software.
Andersen et al. (2022) [14]	✓	△/△/✗	✓	✓	△	✓	✗	✗	Review of non-Markov transition-probability inference; pseudo-observations for regression; limited coverage of informative censoring/IPCW.
Nießl et al. (2023) [18]	✓	✓/✗/✗	✓	✓	△	✗	✗	✗	Theory for non-Markov with left truncation; wild bootstrap; not a broad review.
Aastveit et al. (2023) [19]	△	✗/✓/✗	△	✗	✗	✗	△	✗	Parametric semi-Markov for interval-censored panels; method + `R` package; not a survey of alternatives.
This arti-cle (2025)	✓	✓/✓/✓	✓	✓	✓	✓	✓	✓	Integrated synthesis of non-Markov estimators under LT/IC/InfC; decision rules; ICH E9(R1) alignment; software landscape; bibliometric mapping; simulation on bias/RMSE.

where the coverage indicators are stated as follows: ✓ is covered explicitly; △ is mentioned or partially covered; ✗ is not covered; while the observation-scheme abbreviations are as follows: LT is left truncation; IC is interval or irregular observation; and InfC is informative censoring; with GEE is generalized estimating equations.

The remainder of the article is organized as follows. Section 2 revisits the theoretical and methodological foundations of non-Markov multi-state survival analysis, including estimation under complex censoring. In Section 3, we describe the bibliometric methodology, including search strategy, network construction, and thematic mapping. In Section 4, the empirical results are presented, highlighting publication trends, collaboration networks, and thematic clusters. Section 5 reflects on the methodological implications and practical challenges that emerge from the literature. In Section 6, we conclude with perspectives for future research and practical implementation. The Appendices provide complementary material: the fully annotated `R` code used in the simulation (Appendix A), the exact database search queries employed for study identification (Appendix B), and detailed tabular overviews of the 48 included articles (Appendix C).

2. Theoretical and methodological foundations

This section presents the mathematical toolkit for non-Markov multi-state survival analysis, covering notation, estimation, censoring adjustments, regression frameworks and diagnostic checks.

2.1 *Multi-state framework and notation*

Multi-state models generalize traditional survival analysis by allowing individuals to transition among multiple states over time, rather than experiencing a single terminal event. This framework enables more detailed modeling of disease progression, treatment pathways, or sequences of clinical events, and is particularly suited to settings where intermediate events such as relapse, progression, or adverse reactions influence subsequent risk. While traditional multi-state approaches often assume Markov dynamics, real-world data frequently violate this assumption due to time-dependent effects and history-sensitive transitions. In this subsection, we define the structure and notation of multi-state processes in both Markov and non-Markov settings, establishing the foundation for the estimators and methods discussed throughout the paper. Table 2 summarizes the core symbols used.

To formalize the concepts introduced in Table 2, we now describe the general structure of a multi-state survival process. In this framework, individuals are allowed to move between a finite number of clinically meaningful states over time, capturing key transitions such as disease onset, progression, treatment changes, or death. This setting extends traditional time-to-event models by accommodating intermediate and recurrent events within a unified stochastic process.

Let $\{X(t)\colon t \geq 0\}$ with $X(t) \in S \equiv \{1, \ldots, K\}$ denote a stochastic process representing the state occupied by an individual at time t, taking values in a finite state space S. States are typically numbered so that state K is absorbing (for example, death), while the remaining states may correspond to transient, recurrent, or progressive stages of a disease or treatment pathway.

To formally characterize how individuals move through these states over time, we define, for each ordered pair (i, j) with $i \neq j$, the transition counting process for subject ℓ as

$$N_{ij}^{(\ell)}(t) = \sum_{0<u\leq t} 1_{\{X_\ell(u^-)=i, X_\ell(u)=j\}}, \quad t \geq 0, \tag{1}$$

where 1_B is the indicator function in the set B, which records the number of observed transitions from state i to state j by individual ℓ up to time t. As is standard in this framework, we assume that at most one transition can occur at each time point (the so-called unit-jump assumption, that is, no tied transitions for a given individual at the same instant), ensuring that the process has well-defined increments.

Table 2.: Symbols and definitions used in multi-state survival analysis.

Symbol	Meaning
$X(t)$	State occupied at time t
S	State space
K	Number of distinct states (size of S)
$N_{ij}(t)$	Aggregated counting process for transitions $i \to j$ (sum over individuals)
$Y_i^{(\ell)}(t)$	At-risk indicator for subject ℓ being in state i just before t
$Y_i(t)$	Total number at risk in state i just before t
$\mathcal{H}(t)$	Filtration (full history of the process up to time t)
$\mathcal{H}_\ell(s)$	Observed history of subject ℓ up to time s
$\lambda_{ij}(t\|\mathcal{H}(t^-))$	Conditional transition intensity
$\lambda_{ij}^{\mathrm{M}}(t)$	Markov intensity $\lambda_{ij}(t\|X(t^-)=i)$
$A_{ij}(t)$	Cumulative hazard $\int_0^t \lambda_{ij}(u\|\mathcal{H}(u^-))\mathrm{d}u$
$\widehat{A}_{ij}(t)$	Nelson–Aalen (NA) estimator of $A_{ij}(t)$
$\boldsymbol{A}(t)$	Matrix of cumulative hazards $[A_{ij}(t)]_{i,j=1}^K$
$\widehat{\boldsymbol{P}}(0,t)$	AJ estimator of transition probabilities
$\widehat{P}_{rj}^{\mathrm{LAJ}}(s,t)$	LAJ estimator
$\alpha_{ij}(u)$	Conditional mean intensity $\mathrm{E}(\lambda_{ij}(u \mid \mathcal{H}(u^-))\|X(u^-)=i)$
$y_i(u)$	Occupancy probability $\mathrm{P}(X(u^-)=i)$
$\pi_\ell(t)$	Probability that subject ℓ is uncensored at time t
$\varphi_{n,\ell}(t)$	Influence-function contribution for subject ℓ
Φ	Functional of the cumulative hazard matrix
$\boldsymbol{z}(t)$	Vector of values for time-dependent covariates
$\lambda_{ij}^0(t)$	Baseline intensity for transition $i \to j$
$\boldsymbol{\beta}_{ij}$	Regression coefficients for transition $i \to j$

For individual ℓ, we define the at-risk indicator as

$$Y_i^{(\ell)}(t) = 1_{\{X_\ell(t^-)=i\}}, \tag{2}$$

and for the full cohort, the size of the risk set is given by

$$Y_i(t) = \sum_{\ell=1}^{n} Y_i^{(\ell)}(t), \tag{3}$$

that is, the total number of subjects occupying state i immediately before time t.

Let

$$\mathcal{H}(t) = \sigma\big\{X(u)\colon 0 \le u \le t\big\} \tag{4}$$

denote the history (filtration) of the process up to time t.

The process $\{X(t)\colon t \ge 0\}$ is said to satisfy the Markov property if the distribution of its future state depends only on the present state. Formally, we have that $\mathrm{P}(X(t+h) = j|\mathcal{H}(t), X(t)=i) = \mathrm{P}(X(t+h)=j|X(t)=i)$, for all $h>0$ and $i,j \in S$.

In many biomedical settings this memoryless assumption is violated: the chance of progressing to a more advanced disease stage may depend on the time already spent in the current state or on the particular sequence of earlier states. Consequently, models that permit dependence on the entire past —collectively termed non-Markov multi-state models— are required to capture the complexity of real-world patient trajectories.

We distinguish two standard time scales used in multi-state intensities which are the following: (i) clock-forward (time-since-origin) models index hazards by the absolute study time t since the initial origin; and (ii) clock-reset (sojourn-time) models index hazards by the time spent in the current state, $w = t - T_i$, where T_i is the entry time into state i. Semi-Markov formulations typically adopt the clock-reset scale.

2.2 *Estimators for transition hazards and probabilities*

A central component of these models is the transition intensity function, which quantifies the instantaneous risk of moving from one state to another at a given time. We begin by defining this quantity formally. For any pair of distinct states $i \neq j$, the conditional transition intensity at time t is given by

$$\lambda_{ij}(t \mid \mathcal{H}(t^-)) = \lim_{\Delta \to 0} \frac{\mathrm{P}(X(t+\Delta) = j \mid X(t^-) = i, \mathcal{H}(t^-))}{\Delta}, \tag{5}$$

provided the limit exists. In the expression stated in (5), $\mathcal{H}(t^-)$ denotes the information available just before time t, where $\mathcal{H}(t)$ is the full filtration of the process, as defined in (4). This intensity quantifies the instantaneous probability per unit time of transitioning from state i to state j, given the complete past history up to time t.

When the process satisfies the Markov property, the intensity depends only on the current state and can be written more simply as $\lambda_{ij}^{\mathrm{M}}(t) = \lambda_{ij}(t|X(t^-) = i)$, independent of the prior history. However, in the general non-Markov setting, the intensity may also depend on the duration already spent in the current state or on the full sequence of previously visited states.

To summarize the accumulated risk of transition over time, we define the cumulative transition hazard from state i to state j over the interval $[0, t]$ as

$$A_{ij}(t) = \int_0^t \lambda_{ij}\big(u \mid \mathcal{H}(u^-)\big)\mathrm{d}u. \tag{6}$$

The full matrix of cumulative hazards is denoted by $\boldsymbol{A}(t) = \big(A_{ij}(t)\big)_{i,j=1}^{K}$, where K is the number of states in the process.In Markov models, the diagonal entries are typically set as $A_{ii}(t) = -\sum_{j \neq i} A_{ij}(t)$, mirroring the generator matrix convention in continuous-time Markov chains. Although this constraint is not required in non-Markov models, it is often imposed for analytical convenience. In such cases, the diagonal elements no longer represent cumulative exit hazards but are retained to preserve matrix structure and facilitate computation, especially in product-integral formulations.

Remark 2.1 (Row-sum-zero diagonal) In non-Markov settings, defining $A_{ii}(t) = -\sum_{j \neq i} A_{ij}(t)$ is a bookkeeping device that ensures $\boldsymbol{I} + \mathrm{d}\widehat{\boldsymbol{A}}(u)$ is row-stochastic inside the product integral. The diagonal entries should not be interpreted as estimable cumulative exit hazards per se. Inference should be based on the off-diagonal components and on functionals such as $P_{rj}(s, t)$.

Note that the matrix of cumulative transition hazards continues to offer a foundational representation of transition dynamics. It underpins the estimation of transition probabilities in both Markov and non-Markov multi-state frameworks. In practical applications, these cumulative hazards must be estimated from observed data. A widely used nonparametric approach for this task is based on counting processes and risk sets, and leads to the NA estimator. To formalize this, consider a sample of n independent individuals observed over a fixed follow-up interval $[0, \tau]$, with independent (non-informative) right censoring assumed. Let $N_{ij}^{(\ell)}(t)$ and $Y_i^{(\ell)}(t)$ denote the individual-level transition counting process and at-risk indicator, respectively, as introduced in the expressions formulated in (1) and (2).

Summing over all individuals, we define the aggregated counting and at-risk processes as $N_{ij}(t) = \sum_{\ell=1}^{n} N_{ij}^{(\ell)}(t)$ and $Y_i(t) = \sum_{\ell=1}^{n} Y_i^{(\ell)}(t)$, where $N_{ij}^{(\ell)}(t)$ and $Y_i^{(\ell)}(t)$ are defined in (1) and (2), and the aggregated risk process $Y_i(t)$ is stated in (3).

The cumulative transition hazard $A_{ij}(t)$, introduced in (6), can be estimated nonparametrically using the NA estimator established as

$$\widehat{A}_{ij}(t) = \int_0^t \frac{1}{Y_i(u)} \mathrm{d}N_{ij}(u) = \sum_{u \le t} \frac{\Delta N_{ij}(u)}{Y_i(u)}, \tag{7}$$

where $\Delta N_{ij}(u) = N_{ij}(u) - N_{ij}(u^-)$ denotes the increment of the counting process at time u; the sum is taken over event times u such that $Y_i(u) > 0$. In practical implementations, the denominator is often stabilized by replacing $Y_i(u)$ with $Y_i(u) \vee 1 = \max\{Y_i(u), 1\}$, to avoid division by zero when the risk set is temporarily empty. This does not affect consistency under standard regularity conditions but improves numerical stability in finite samples.

Under suitable asymptotic conditions —specifically, assuming that the at-risk process $Y_i(u)$ remains strictly positive at all observed event times and that the number of individuals n tends to infinity— the NA estimator is consistent and asymptotically normal. In particular, the scaled estimation error satisfies $\sqrt{n}(\widehat{A}_{ij}(t) - A_{ij}(t))$, where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution,and the asymptotic variance is $V_{ij}(t) = \int_0^t (\alpha_{ij}(u)/y_i(u))\mathrm{d}u$, with $\alpha_{ij}(u) = \mathrm{E}(\lambda_{ij}(u \mid \mathcal{H}(u^-))\big|X(u^-) = i)$ being the conditional mean transition intensity out of state i, and $y_i(u) = \mathrm{P}(X(u^-) = i)$.

These results provide the theoretical foundation for constructing pointwise confidence intervals for $A_{ij}(t)$, either via plug-in variance estimates (for example, Greenwood-type formulas) or nonparametric bootstrap procedures based on resampling individuals. Importantly, the NA estimator also serves as a building block for estimating transition probabilities in both Markov and non-Markov multi-state models.

In the Markov setting, the transition probability matrix from time 0 to time t, denoted $\widehat{\boldsymbol{P}}(0,t)$, can be estimated using the AJ estimator defined as

$$\widehat{\boldsymbol{P}}(0,t) = \mathrm{Prod}_{u \le t}\big(\boldsymbol{I} + \mathrm{d}\widehat{\boldsymbol{A}}(u)\big), \tag{8}$$

where $\mathrm{Prod}_{u \le t}\big(\boldsymbol{I} + \mathrm{d}\widehat{\boldsymbol{A}}(u)\big)$ denotes the product integral, the multiplicative analogue of the Riemann integral for matrix-valued right-continuous-with-left-limits (càdlàg) processes. It accumulates infinitesimal transition updates into a row-stochastic estimate of $\boldsymbol{P}(0,t)$.

The diagonal entries of $\widehat{\boldsymbol{A}}(u)$ are defined as $\widehat{A}_{ii}(u) = -\sum_{j \ne i} \widehat{A}_{ij}(u)$, ensuring that each factor $\boldsymbol{I} + \mathrm{d}\widehat{\boldsymbol{A}}(u)$ is row-stochastic so that rows of $\widehat{\boldsymbol{P}}(0,t)$ sum to one.

Remark 2.2 (AJ under non-Markov) Even if the process is non-Markov, and provided censoring is non-informative, the expression presented in (8) consistently estimates $\boldsymbol{P}(0,t)$ for the Markovized process with generator $\boldsymbol{\alpha}(t) = [\alpha_{ij}(t)]$, where $\alpha_{ij}(t) = \mathrm{E}(\lambda_{ij}(t|\mathcal{H}(t^-))|X(t^-) = i)$ [3, 12, 14, 15].

To orient practical choices after the expression formulated in (8), the following summary outlines when each family of estimators is appropriate, given the data structure and the likely presence of history dependence:

The following is a practical guide to selecting estimators in multi-state analyses:

- Baseline use —Use the AJ estimator defined in (8) for $\boldsymbol{P}(0,t)$ when censoring is plausibly non-informative. Note that AJ remains a defensible baseline even under non-Markov dynamics because it targets the transition matrix built from conditional mean intensities. For conditional probabilities $P_{rj}(s,t)$ with $s > 0$, prefer LAJ or a non-Markov plug-in unless the Markov property is defensible.
- Indications of history dependence —Prefer a LAJ analysis (conditioning at a fixed s) or a non-Markov plug-in estimator of $P_{rj}(s,t)$ [11, 12, 14, 20], when hazards vary with time already spent in the current state or with the timing/sequence of prior events.
- Regression on transition or state-occupation probabilities —Utilize pseudo-observations combined with generalized estimating equations (GEE) to obtain population-level covariate effects without full likelihood specification; see Section 2.3.
- Informative censoring —Incorporate IPCW and combine it with the chosen estimator (baseline AJ, LAJ, or plug-in), checking the censoring model specification, when dropout depends on past history.
- Semi-Markov structure —Formulate intensities with sojourn-time dependence (clock-reset or clock-forward), which aligns the model with clinical path dependence, if time-in-state is mechanistically relevant.
- Complex designs and hierarchical data —Consider Bayesian hierarchical formulations for interval censoring, clustering, or joint longitudinal–survival structures, where likelihood functions are computationally heavy saddlepoint approximations may offer computational acceleration.
- Diagnostics and design considerations —Contrast AJ and landmark estimates, stratify by sojourn time, and perform sensitivity analyses to the visit process and weights, before committing to a model, ensuring risk-set construction correctly handles delayed entry and irregular observation.

For quick method selection across data structures and estimands, Table 3 summarizes the main estimators and modeling families reviewed, with their applicability conditions, strengths, and caveats.

Table 3.: Comparison of methods/estimators for multi-state survival with history dependence and complex censoring.

Method/ estimator	Core idea and estimand	Applicability conditions	Strengths	Limitations/cautions
AJ	Product integral of NA to estimate transition probabilities $P(s,t)$ under a (time-inhomogeneous) Markov process.	Independent right-censoring (or censoring handled by design); correct risk-set construction; left truncation handled via delayed entry; interval censoring needs dedicated extensions.	Nonparametric; well-studied asymptotics; widely implemented; natural baseline comparator.	Biased under history dependence (non-Markov); not robust to informative censoring without weighting; limited for interval-censored panels without specialized methods.

Continued on next page

Table 3 (continued).

Method/ estimator	Core idea and estimand	Applicability conditions	Strengths	Limitations/cautions
LAJ	Condition on state at landmark time s and apply AJ on $(s,t]$ within that subcohort; dynamic $P_{rj}(s,t)$.	Non-Markov settings; independent censoring conditional on the landmark (or IPCW); adequate at-risk counts per landmark; left truncation allowed with proper risk sets.	Handles history dependence locally; simple; interpretable dynamic predictions; aligns with clinical follow-up.	Efficiency loss with many landmarks; risk of selection/immortal-time bias if landmarking or risk-set construction is mis-specified; still sensitive to informative censoring unless IPCW is used.
Non-Markov nonparametric $P(s,t)$ (such as illness–death plug-in)	Direct non-Markov estimators of $P_{rj}(s,t)$ via conditional cumulative incidence/plug-in functionals of hazards.	Progressive and general graphs; right-censoring; extendable to left truncation and (with care) interval censoring; IPCW possible.	Consistent under history dependence; transparent assumptions; good for progressive models.	Implementation more involved; variance often via bootstrap; care with sparse paths and long horizons.
Semi-Markov intensity models	Transition intensities depend on sojourn time (clock-reset/ clock-forward); $h_{ij}(t,w)$ with time-in-state w.	History dependence driven by time-in-state; typically right-censoring; can include covariates/frailty.	Flexible for path dependence; clinically interpretable (such as post-progression hazards).	Model mis-specification of time-in-state effects can bias results; heavier computation; diagnostics needed.
Illness–death/semi-competing risks	Jointly model nonterminal event (such as progression) and death; typically non-Markov.	Progressive structures; right-censoring; extensions exist for interval censoring/left truncation.	Natural estimands (PFS/OS); explicit dependence between paths; policy-relevant.	Identifiability and sensitivity to modeling choices; careful censoring handling required.
IPCW (inverse-probability of censoring weighting)	Reweight contributions by $\pi^{-1}(t)$ to correct for informative/dropout censoring in hazards or $P(s,t)$.	Censoring depends on history/covariates but is modelable; correct weight model; positivity.	Mitigates bias from informative censoring; integrates with AJ/LAJ/plug-in frameworks.	Sensitive to misspecification; extreme weights inflate variance; diagnostics essential.
Pseudo-observations GEE (for $P_{rj}(s,t)$ or state occupation)	Construct pseudo-observations for target functionals and regress via GEE for marginal covariate effects.	Population-level effects; clustering accommodated; censoring handled in construction.	Direct inference on probabilities; marginal interpretation; avoids full likelihood.	Choice of time grid/functional matters; small-sample bias possible; requires robust SEs.

Continued on next page

Table 3 (continued).

Method/ estimator	Core idea and estimand	Applicability conditions	Strengths	Limitations/cautions
Bayesian hierarchical/MCMC (incl. semi-competing risks)	Full probabilistic models for transitions/histories with priors, random effects, and joint components.	Complex data (history dependence, clustering, longitudinal biomarkers); prior elicitation.	Joint uncertainty; shrinkage; handles missingness; posterior predictive checks.	Compute-intensive; convergence diagnostics; scalability to very large EHR cohorts limited.
Saddle-point/ flowgraph and accelerators	Approximate transforms/path sums to accelerate likelihoods or obtain fast summaries.	Large state spaces; rare events; need fast approximate inference.	Big speed-ups; tractable summaries for complex paths.	Specialist assumptions; niche tooling; validation versusexact/MC needed.
Interval-censoring/ left-truncation extensions	Penalized likelihood or nonparametric maximum likelihood estimation (NPMLE)/Turnbull-type estimators tailored to observation schemes.	Visit-scheduled cohorts; delayed entry; registry data with irregular observation.	Uses all data correctly; reduces bias from visit processes.	Implementation complexity; computational burden; careful risk-set construction crucial.
Bootstrap/ subsampling	Resampling for SEs/CIs or rare-event scaling to big data.	Analytic variances hard; massive cohorts; individual participant data (IPD) aggregation.	Generic; cross-estimator; supports robustness checks.	Heavy in large samples; must respect clustering/history in resampling.

These desirable Markov properties hinge on the Markov assumption for conditional targets. When this assumption does not hold—so that future evolution depends on the past beyond the current state—applying the expression stated in (8) without conditioning at a landmark generally yields biased estimates of conditional transition probabilities. By contrast, for unconditional probabilities from time 0 and under independent censoring, the formulation presented in (8) remains consistent for the Markovized process driven by conditional mean intensities. In such non-Markov settings, two classes of estimators are used for conditional targets $\boldsymbol{P}(s,t)$: (i) LAJ and (ii) non-Markov plug-in estimators based on conditional survival/cumulative incidence.

The first approach, often referred to as the LAJ estimator, is based on stratifying the analysis by a fixed landmark time s and conditioning on the subset of individuals known to occupy a specific state r at that time. Within this subcohort, the standard AJ estimator is applied to the interval $(s,t]$, yielding

$$\widehat{P}_{rj}^{\mathrm{LAJ}}(s,t) = \left[\mathrm{Prod}_{u\in(s,t]}\big(\boldsymbol{I} + \mathrm{d}\widehat{\boldsymbol{A}}^{(s)}(u)\big)\right]_{rj}. \tag{9}$$

where $\widehat{\boldsymbol{A}}^{(s)}(u)$ denotes the matrix of NA estimators constructed using only those individuals in state r at time s. The at-risk set used in this computation is given by

$$Y_i^{(s)}(u) = \sum_{\ell=1}^{n} 1_{\{X_\ell(s)=r, X_\ell(u^-)=i\}}, \tag{10}$$

which ensures that the estimation reflects the subpopulation defined by the landmark condition. Structurally, the risk set $Y_i^{(s)}(u)$ stated in (10) mirrors the general at-risk indicator $Y_i^{(\ell)}(t)$ introduced in (2), but is restricted to individuals known to be in state r at the landmark time s.

This estimator is consistent provided that censoring is independent (or correctly modeled via IPCW) conditional on the landmark state $X(s) = r$ and the observed covariates, with risk sets constructed to account for delayed entry. A detailed discussion and theoretical justification can be found in [12].

An alternative strategy involves estimating transition probabilities directly from the individual's observed history, without assuming the Markov property. In this framework, one seeks to approximate the conditional probability given by $\mathrm{P}(X(t) = j|X(s) = r, \mathcal{H}(s))$, where $\mathcal{H}(s)$ denotes the full event and covariate history up to time s, as defined in (4).

For $j = r$, in general (non-Markov) there is indeed no simple survival-ratio representation. A correct identity is given by

$$P_{rr}(s,t) = \mathrm{E}\left(\exp\left(-\int_s^t \lambda_{r\cdot}\big(u \mid \mathcal{H}(u^-)\big)\mathrm{d}u\right)\middle| X(s) = r\right),$$

where $\lambda_{r\cdot}(u) = \sum_{k\neq r}\lambda_{rk}(u|\mathcal{H}(u^-))$. This does not in general reduce to $\exp(-(A_{r\cdot}(t) - A_{r\cdot}(s)))$ with $A_{r\cdot}(u) = \sum_{k\neq r}\int_0^u \alpha_{rk}(v)\mathrm{d}v$ and $\alpha_{rk}(v) = \mathrm{E}(\lambda_{rk}(v \mid \mathcal{H}(v^-))|X(v^-) = r)$, where by the Jensen inequality, $\mathrm{E}(\exp(-Z)) \neq \exp(-\mathrm{E}(Z))$ in general. In practice, we therefore estimate $P_{rr}(s,t)$ either via LAJ —expression given in (9)— or via a non-Markov plug-in based on conditional cause-specific hazard increments computed within the landmark subcohort.

A consistent plug-in under history dependence uses conditional increments taken within the landmark subcohort. Let $\alpha_{rj}^{(s)}(u) = \mathrm{E}(\lambda_{rj}(u|\mathcal{H}(u^-))|X(s) = r, X(u^-) = r)$, where $A_{rj}^{(s)}(u) = \int_{(s,u]}\alpha_{rj}^{(s)}(v)\mathrm{d}v$. Then, we get $\widehat{P}_{rj}(s,t) = \int_{(s,t]}\widehat{P}_{rr}(s,u^-)\mathrm{d}\widehat{A}_{rj}^{(s)}(u)$, where the conditional increment $\mathrm{d}\widehat{A}_{rj}^{(s)}(u)$ is computed using the landmark risk set $Y_r^{(s)}(u) = \sum_{\ell=1}^n 1_{\{X_\ell(s)=r, X_\ell(u^-)=r\}}$ from the expression established in (10).

If censoring depends on past history, employ IPCW with Horvitz–Thompson weighting [21]in both the numerator and the denominator of the NA increment stated as $\mathrm{d}\widehat{A}_{rj}^{(s,\mathrm{IPCW})}(u) = \sum_{\ell=1}^n w_\ell(u)\mathrm{d}N_{rj}^{(s,\ell)}(u)/\sum_{\ell=1}^n w_\ell(u)Y_r^{(s,\ell)}(u)$, where $w_\ell(u) = (\widehat{\pi}_\ell(u))^{-1}$, with $\pi_\ell(u) = \mathrm{P}(C_\ell \geq u|\mathcal{H}_\ell(u))$. In practice, stabilized or truncated weights help to control variability; validity requires correct specification of the censoring model and a positivity condition (non-zero probability of remaining uncensored across covariate histories).

This yields a non-Markov estimator aligned with the target $P_{rj}(s,t)$. These non-Markov estimators allow for more general forms of history dependence and have demonstrated good performance under complex observation schemes, including left-truncation and informative censoring. For comprehensive comparisons and applied illustrations, see [11, 14].

Both landmark-based and history-dependent plug-in estimators provide flexible and interpretable tools for estimating transition probabilities when the Markov assumption does not hold. Their relative performance depends on the structure of the process, the censoring pattern, and the specific transition quantities of interest.

Beyond point estimation, it is needed to quantify the uncertainty associated with estimated hazards or transition probabilities —for example, to construct confidence intervals or conduct hypothesis testing. A central step in such inference is the estimation of the variance associated with a given estimator, which we denote generically by $\widehat{\theta}_n(t)$. Suppose that this estimator can be expressed as a smooth functional $\Phi(\widehat{\boldsymbol{A}}(t))$ of the estimated cumulative hazard matrix $\widehat{\boldsymbol{A}}(t) = [\widehat{A}_{ij}(t)]_{i,j}$, as is the case for both NA and AJ estimators.

Using a first-order Taylor expansion around the true cumulative hazard matrix, one obtains the influence function (IF) representation given by $\varphi_{n,\ell}(t) = \sum_{i\neq j}\int_0^t(\partial\Phi/\partial A_{ij}(u))\mathrm{d}M_{ij}^{(\ell)}(u)/Y_i(u)$, where $\partial\Phi/\partial A_{ij}(u)$ is evaluated at the true cumulative hazard $\boldsymbol{A}(t)$, and where $M_{ij}^{(\ell)}(u) = N_{ij}^{(\ell)}(u) - \int_0^u \lambda_{ij}(s|\mathcal{H}_\ell(s^-))Y_i^{(\ell)}(s)\mathrm{d}s$ is the martingale associated with the counting process for subject ℓ. Here, $M_{ij}^{(\ell)}(u)$ is the individual-level martingale associated with the counting process for subject ℓ. Moreover, $Y_i(u) = \sum_{\ell=1}^n Y_i^{(\ell)}(u)$ denotes the total number of individuals at risk in state i just prior to time u. This representation captures the asymptotic contribution of each subject to the variability of the estimator. However, note that this IF approach assumes that Φ is a smooth (that is, differentiable) functional of the cumulative hazard matrix. In applications involving non-smooth functionals —such as transition time quantiles, survival percentiles, or stepwise classification rules— the delta method may not apply directly. In such cases, alternative techniques such as analytic linearization, jackknife variance estimation, or nonparametric bootstrap procedures are recommended.

Define the (unscaled) IF by means of $\sqrt{n}(\widehat{\theta}_n(t) - \theta(t)) = n^{-1/2}\sum_{\ell=1}^n \mathrm{IF}_\ell(t) + o_p(1)$. Equivalently, set $\varphi_{n,\ell}(t) = \mathrm{IF}_\ell(t)/\sqrt{n}$. Then, we have $\widehat{\mathrm{Var}}(\widehat{\theta}_n(t)) = (1/n^2)\sum_{\ell=1}^n \widehat{\mathrm{IF}}_\ell^2(t) = (1/n)\sum_{\ell=1}^n \widehat{\varphi}_{n,\ell}^2(t)$. In what follows we adopt the latter convention and treat $\varphi_{n,\ell}(t)$ as already scaled by $1/\sqrt{n}$. This yields the familiar Greenwood–type estimator in the multistate setting given by $\widehat{\mathrm{Var}}(\widehat{\theta}_n(t)) = (1/n)\sum_{\ell=1}^n \widehat{\varphi}_{n,\ell}^2(t)$. When subjects are clustered (for example, centers, families), individual-level independence may fail. Let clusters be indexed by $g \in \{1,\ldots,G\}$ and define the cluster influence contribution $\psi_g(t) = \sum_{\ell\in\mathcal{C}_g}\widehat{\varphi}_{n,\ell}(t)$. A cluster-robust (sandwich) variance is then stated as $\widehat{\mathrm{Var}}_{\mathrm{clust}}(\widehat{\theta}_n(t)) = (1/n)\sum_{g=1}^G \psi_g^2(t)$, optionally with small-sample corrections, such as factors $G/(G-1)$ and $n/(n-1)$. As an alternative, a cluster (block) bootstrap that resamples entire clusters can be used to account for within-cluster dependence. Provided differentiability holds, these forms apply to a wide class of functionals Φ, including those for transition probabilities and cumulative hazards. Provided differentiability holds, this general form applies to a wide class of functionals Φ, including those involved in the estimation of transition probabilities and cumulative hazards.

In many cases, such as the NA estimator for $A_{ij}(t)$ or the AJ estimator for $\boldsymbol{P}(s,t)$, explicit expressions for the variance are available and can be derived analytically using martingale theory and the delta method. When such closed-form expressions are not tractable, a nonparametric bootstrap procedure—based on resampling individuals with replacement—provides a practical and reliable alternative. Because the bootstrap operates at the subject level, it preserves the dependence structure across transitions and allows consistent estimation of variances in both Markov and non-Markov settings.

So far, our discussion has assumed that individuals are observed from a common time origin and that transitions are recorded exactly. However, in many real-world biomedical applications, more complex observation schemes arise—such as delayed entry, interval-censored transitions, and informative censoring. This subsection summarizes how standard estimators can be adapted to these scenarios.

In the case of left truncation (delayed entry), subjects contribute only after an entry time $L_\ell > 0$. Under independent truncation, replace the at-risk indicator given in (2) with $Y_i^{(\ell)}(t)\,1_{\{L_\ell\leq t\}}$. If all subjects enter through state 1, this suffices. If some enter directly into $k \neq 1$, initialize both $Y_i^{(\ell)}(t)$ and the relevant counting processes $N_{kj}^{(\ell)}(t)$ at $t = L_\ell$ from the appropriate starting state. With these modifications, the NA and AJ estimators presented in (7)–(8) remain valid.

In the case of interval censoring, state transitions are not observed exactly but only known to occur within a time interval $(L_\ell, R_\ell]$, as is common in studies with periodic assessments. One option is to apply the Turnbull self-consistency algorithm for NPMLE under interval censoring [22]. Alternatively, midpoint imputation within each interval may be used, but this is merely a pragmatic shortcut; it is generally biased and not consistent except under vanishing-interval asymptotics. For consistent nonparametric inference under interval censoring, use the Turnbull NPMLE or penalized-likelihood approaches; see [23].

In the case of informative censoring, when the censoring time C_ℓ depends on the individual's past history or event trajectory, naive estimators may be biased. A general correction strategy is IPCW. Specifically, each individual's contribution is weighted by the inverse of the estimated survival function given by $\pi_\ell(t) = \mathrm{P}(C_\ell \geq t|\mathcal{H}_\ell(t))$, reflecting the probability of being uncensored at time t. This yields IPCW-adjusted versions of the NA and related estimators that are unbiased under a correctly specified model for the censoring distribution and a positivity condition. The IPCW approach in non-Markov settings is discussed in detail in [24].

These extensions enable the general framework of cumulative hazard and transition-probability estimation to be applied in settings that reflect the complexities of clinical data collection, including irregular observation times and incomplete follow-up.

As an orientation aid, Figure 1 summarizes the decision flow. Design and observation-scheme features inform the choice of model family, where the top arrows represent a gradient of history dependence rather than strict set inclusion. Based on this, one selects an appropriate estimator family —AJ for unconditional targets $P(0,t)$ and LAJ/plug-in for conditional targets $P_{rj}(s,t)$— and, if required, adds cross-cutting computational adjustments (such as IPCW or bootstrap). Before reporting transition and state-occupation probabilities and dynamic predictions, diagnostics are performed; these may in turn motivate revisiting both the model family and the estimator.

2.3 *Regression, diagnostics, and prediction in non-Markov models*

Multi-state models can be enriched by incorporating covariate information that affects transition intensities. A common and flexible formulation expresses the intensity for a transition from state i to state j at time t as $\lambda_{ij}(t|\mathcal{H}(t^-)) = \lambda^0_{ij}(t)\exp(\boldsymbol{\beta}_{ij}^\top \boldsymbol{z}(t))$, where $\lambda^0_{ij}(t)$ is a baseline intensity function, $\boldsymbol{\beta}_{ij}$ is a vector of regression coefficients, and $\boldsymbol{z}(t)$ denotes a vector of values for covariates that may include time-dependent variables or features derived from the subject's history. This structure accommodates both Markov and semi-Markov dynamics, depending on how the history is encoded into the covariate process $\boldsymbol{z}(t)$.

Estimation in this framework typically proceeds via partial likelihood, extending the Cox model to stratified transition types. Each transition $i \to j$ is treated as a separate stratum with its own risk set, and inference follows standard Cox procedures. This approach, originally formalized in [25], allows flexible modeling of event-specific hazards in multi-state settings with time-dependent covariates. For valid inference, it is assumed that censoring is conditionally independent given covariates; if unobserved heterogeneity (for example, frailties shared across transitions) is present, this assumption may be violated, potentially leading to biased estimates of $\boldsymbol{\beta}_{ij}$. In such cases, frailty-adjusted models or IPW may be required.

In cases where interest lies in transition probabilities $P_{rj}(s,t)$ rather than intensities, pseudo-observation methods and GEE provide viable alternatives. These allow covariate effects to be incorporated even when full likelihood-based specification is intractable or undesirable. Importantly, such methods remain valid under non-Markov assumptions; see [14] for a recent overview.

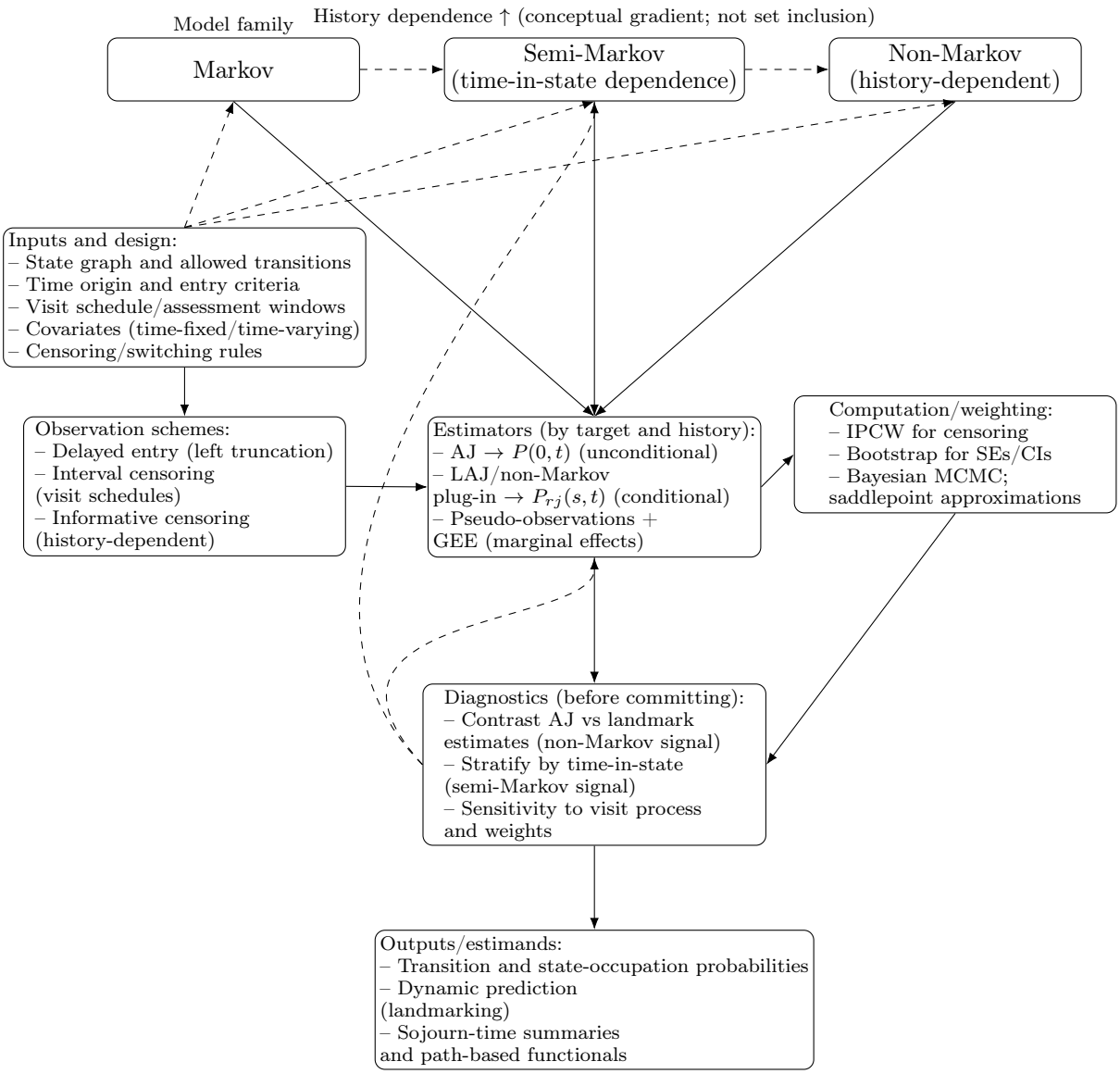

Figure 1.: Conceptual map of the multi-state analysis workflow.

Before choosing between Markov, semi-Markov, or more general history-dependent models, it is often helpful to assess whether the Markov assumption is empirically supported by the data. One diagnostic strategy, originally proposed in [15], involves stratifying individuals currently in state i according to the time already spent in that state. Estimated cumulative hazards $\widehat{A}_{ij}(t)$ are then compared across strata. If the curves are approximately parallel, this suggests that transition intensities do not depend substantially on sojourn time, lending empirical support to the Markov assumption.

More formal comparisons can be made between transition probabilities estimated via the standard AJ estimator given in (8), which assumes the Markov property, and those obtained via the LAJ estimator stated in (9), which does not. Substantial discrepancies between the two may signal violations of the Markov assumption. In such cases, hypothesis testing based on wild bootstrap resampling, as described in [13], offers a practical tool for significance assessment. However, this approach relies on independence between subjects and requires a moderate sample size. When these assumptions are not met—such as in cluster-correlated data or in smaller cohorts —pseudo-observation-based diagnostic methods may offer a more robust alternative [14].

These model diagnostics are especially relevant in clinical applications that demand individualized prediction. A common goal is to estimate the probability that a subject —observed up to time s, with covariate history and event trajectory summarized in $\mathcal{H}_\ell(s)$— will occupy a given state j at a future time $s+w$. This conditional transition probability, often termed a dynamic prediction, can be estimated using nonparametric methods such as the LAJ estimator, applied to individuals known to be in a specific state at the landmark time. Alternatively, landmark-specific Cox models may be used, allowing covariates measured at time s to influence subsequent transition risks. These models have been shown to provide reliable and interpretable predictions even under violations of the Markov assumption.

In more complex settings, particularly those involving longitudinal biomarkers, recurrent assessments, or latent health states, joint modeling approaches offer a principled alternative. For example, dynamic prediction can be performed using joint frailty models or Bayesian hierarchical models that incorporate time-varying covariates and subject-specific random effects. Tensor regression frameworks, which summarize multidimensional biomarker trajectories, have also shown promise for producing personalized survival predictions [26, 27]. Such tools expand the applicability of multi-state models in precision medicine, enabling more nuanced risk stratification and treatment planning.

Each of these frameworks provides tools for personalized prognosis in complex disease processes, adapting to both the clinical context and the data structure. The choice among them depends on the desired balance between interpretability, model complexity, and computational feasibility.

Having discussed estimation, weighting, and diagnostics for non-Markov multi-state analysis, we next provide a compact numerical illustration to make the implications concrete; see Section 2.4.

2.4 *Simulation under semi-Markov dynamics and history-dependent censoring*

To illustrate how (i) history dependence in the post-progression hazard and (ii) history-dependent censoring affect transition-probability estimation, we ran a focused simulation comparing the AJ estimator presented in (8) with its LAJ variant. The estimand is the death probability $P_{13}(s,t)$ in a progressive illness–death graph ($1 \to 2 \to 3$ and $1 \to 3$).

Note that, under non-informative censoring, baseline AJ is consistent for the unconditional target $P_{13}(0,t)$ even when the post-progression hazard is semi-Markov; here we deliberately induce history-dependent censoring to study its impact on AJ/LAJ.

We simulate $n = 500$ individuals per replicate from a three-state process with constant hazards out of state 1 stated as $h_{12} = 0.15, h_{13} = 0.05$, and a semi-Markov death intensity after progression that depends on the sojourn time w in state 2 and on the time of progression T_P presented as $h_{23}(w, T_P) = c(T_P)(w+1)^{0.3}, c(T_P) = 0.08\exp(0.4 \times 1_{\{T_P>6\}})$. Censoring is history-dependent: the censoring hazard is $\lambda_C(t) = 0.02\exp(0.6 \times 1_{\{t>T_P\}})$ (that is, it increases once progression occurs). Administrative end is at $\tau_{\max} = 24$. We evaluate $t \in \{6, 12, 18, 24\}$ and landmark times $s \in \{6, 12\}$. We run 200 MC replicates.

AJ (baseline) estimates $P_{13}(0,t)$ using the standard product-integral of NA increments (no IPCW —deliberately omitted to isolate the impact of history-dependent censoring), while LAJ estimates $P_{13}(s,t)$ by restricting to the subcohort with $X(s) = 1$ and re-originating time at s before applying AJ on $(s,t]$. Both are computed with `mstate` (`coxph` with `strata(trans)` followed by `msfit`/`probtrans`) and risk sets are constructed to ensure consistent exit times for the $1 \to 2$ and $1 \to 3$ transitions. Truth values Truth(s,t) are approximated via a large, uncensored Monte Carlo (MC) sample ($n = 200{,}000$). For $s = 0$ we compute $P_{13}(0,t)$ unconditionally. For $s > 0$, we condition on the landmark subcohort with $X(s) = 1$ and evaluate $P_{13}(s,t)$ within that subpopulation. For each method and (s,t) we report the across-replicate mean estimate, bias = mean − truth, and RMSE.

Table 4.: Simulation results for $P_{13}(s,t)$ under semi-Markov dynamics and history-dependent censoring.

Method	s	t	n_{reps}	Mean	Truth	Bias	RMSE
AJ (baseline)	0	6	200	0.394	0.340	0.055	0.059
AJ (baseline)	0	12	200	0.736	0.665	0.071	0.074
AJ (baseline)	0	18	200	0.906	0.865	0.041	0.044
AJ (baseline)	0	24	200	0.971	0.953	0.018	0.020
LAJ ($s = 6$)	6	12	200	0.441	0.393	0.048	0.066
LAJ ($s = 6$)	6	18	200	0.775	0.733	0.042	0.057
LAJ ($s = 6$)	6	24	200	0.923	0.902	0.021	0.033
LAJ ($s = 12$)	12	18	200	0.437	0.399	0.037	0.089
LAJ ($s = 12$)	12	24	200	0.774	0.739	0.035	0.086

where each cell aggregates 200 replicates ($n = 500$); truth is computed from a large-n MC sample without censoring; and for $s > 0$, Truth(s,t) is computed conditionally on $X(s) = 1$ in that large, uncensored MC sample.

Across all (s,t), the estimators overestimated $P_{13}(s,t)$ (positive bias). For the unconditional target $P_{13}(0,t)$, AJ bias declines with t (from 0.055 at $t = 6$ to 0.018 at $t = 24$), reflecting diminishing headroom near the administrative horizon. Conditioning at a landmark reduces absolute bias for $P_{13}(s,t)$ —such as LAJ at $s = 6$ yields biases of 0.048, 0.042, and 0.021 at $t = 12, 18, 24$— but residual bias persists under history-dependent censoring. The later landmark ($s = 12$) shows similar mean bias (such as 0.037 at $t = 18$) but noticeably larger RMSE (≈ 0.0875), consistent with a smaller risk set and shorter prediction window. Mechanistically, once progression occurs the censoring hazard increases, so individuals with a lower post-progression death hazard —such as early progressors with smaller $c(T_P)$— spend more time exposed to censoring and are removed from observation more often. This preferential censoring of lower-risk paths leaves a higher-risk mix among those observed, inflating the empirical $2 \to 3$ intensity and thus $P_{13}(s,t)$. Landmark conditioning reduces, but does not eliminate, this selection.

Note that landmarking is a simple effective way to mitigate non-Markov bias, when hazards depend on time since an intermediate event; while if censoring depends on past history, LAJ estimates remain biased unless censoring is modeled (such as IPCW); and later landmarks trade bias for variance due to smaller risk sets. This aligns with the decision rules outlined earlier (Section 2) and motivates the incorporation of weights/diagnostics in applied workflows. The design is intentionally minimal (no covariates, fixed administrative end, single progressive graph). We did not include IPCW, pseudo-observations, or Bayesian formulations here; the goal is pedagogical —isolating the effects of semi-Markov dynamics and history-dependent censoring on AJ-type estimators. Fully annotated `R` code that generates Table 4 is provided in Appendix A.

3. Bibliometric Methodology

This section outlines the procedures used to identify, select, and quantitatively analyze the 48 articles included in the review. Bibliometric mapping has been increasingly applied in biomedical and clinical research to chart research landscapes and collaboration patterns [28], but to our knowledge no prior work has combined a technical synthesis of non-Markov multi-state methods with a bibliometric analysis of the corresponding methodological field. Therefore, we describe the search strategy, construction of co-occurrence networks, and application of graph-based techniques for clustering and thematic mapping.

3.1 *Search strategy and inclusion criteria*

To provide a structured overview of recent advances in non-Markov multi-state survival analysis under complex censoring mechanisms, we conducted a systematic literature review following the PRISMA 2020 guidelines [29]. The goal was to identify peer-reviewed journal articles that contribute substantively—either methodologically or through applied modeling —to multi-state survival frameworks that relax the Markov assumption and incorporate non-standard censoring features.

Database searches were performed in Web of Science and Scopus, covering all publication years up to 11 June 2025. Searches were run on titles, abstracts, and author keywords, restricted to English-language records. Duplicate records across databases were resolved at the digital object identifier (DOI) level; when metadata differed, we retained the entry with the richer bibliographic fields. No review protocol was preregistered. The search queries combined terms from three conceptual domains, with expressions within each block joined by `OR`, and the blocks themselves connected by `AND`. These three conceptual domains are the following:

(i) Time-to-event terminology: "survival analysis" OR "time-to-event" OR "time to event".
(ii) Modeling and estimation: "multi-state" OR multi-state OR "non-Markov" OR "landmark approach" OR "landmark model" OR "transition probabilit*" OR pseudo-observ* OR GEE OR IPCW OR "semi-Markov" OR "illness–death" OR "semi-competing risk*" OR subsampling.
(iii) Observation and censoring schemes: censor* OR "interval censor*" OR "left trunc*" OR "right censor*" OR irregular OR "panel data".

These terms were refined through preliminary searches based on foundational studies [6, 13], ensuring adequate coverage of relevant approaches, particularly those involving history-dependent transitions and estimation under complex observation schemes. The full Boolean syntax is provided in Appendix B.

Initial queries retrieved 374 records. To ensure the relevance and quality of included studies, we applied the following filtering criteria:

- Inclusion of peer-reviewed items (journal articles and peer-reviewed conference proceedings).
- Exclusion of records lacking core metadata (title, abstract, or DOI), consistent with the pre-specified English-language scope.
- Removal of any residual duplicates after DOI harmonization across databases.

Eligibility was then assessed in two stages. First, titles and abstracts were screened independently by two reviewers using three predefined criteria: (i) use of a multi-state or recurrent-event framework; (ii) explicit relaxation or empirical assessment of the Markov assumption; (iii) modeling of at least one form of complex censoring, including left truncation, interval censoring, or informative dropout. Non-peer-reviewed documents (for example, theses, conference abstracts) and purely narrative reviews were excluded. Discrepancies were resolved by discussion.

The second stage involved full-text assessment of the remaining articles to extract methodological features, censoring structures addressed, and statistical techniques employed. No further exclusions occurred at this point.

Following this process, 48 articles were retained for inclusion, corresponding to 18.2% of the screened set. Figure 2 summarizes the selection workflow using the PRISMA-2020 format. The PRISMA diagram in Figure 2 summarizes the selection workflow using the PRISMA-2020 format. We identified 374 records across Web of Science and Scopus, removed 98 duplicates and 12 records with incomplete metadata (missing DOI, title, or abstract), and screened 264 titles/abstracts.

Of the 264 articles, 216 were excluded for not meeting pre-specified criteria (no multi-state/recurrent-event framework, Markov-only analyses, absence of complex censoring; non–peer-reviewed items; or purely narrative reviews). All 48 records retrieved at full text were eligible and entered the bibliometric synthesis.

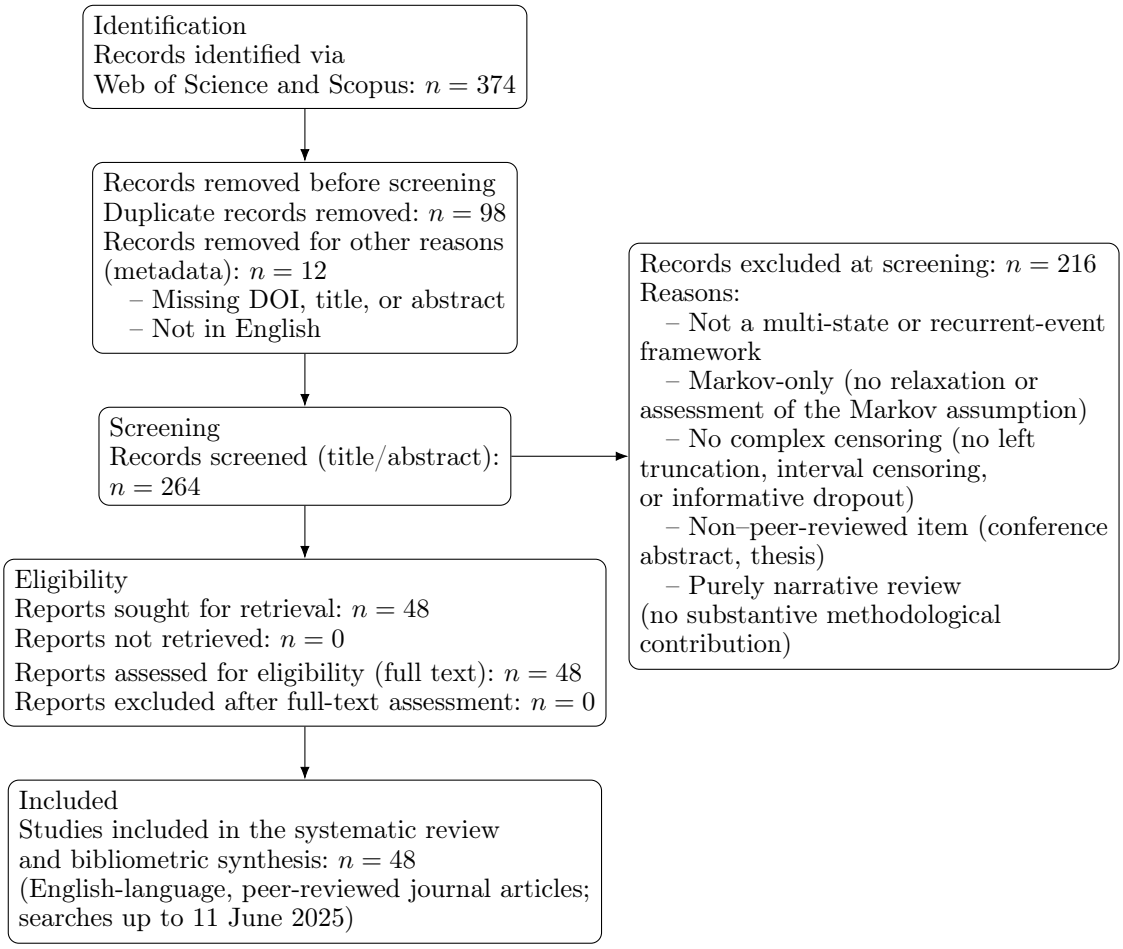

Figure 2.: PRISMA 2020 flow diagram with counts and reasons per stage.

With the final corpus established, the next step was to convert their bibliographic metadata into quantitative structures amenable to network analysis. In particular, we built co-occurrence matrices for authors, keywords, and country affiliations and then normalized those matrices to obtain weighted graphs. The rationale and formulas behind this transformation are detailed below.

3.2 *Construction and normalization of co-occurrence matrices*

Bibliometric network analysis begins with the construction of a co-occurrence matrix that quantifies how often pairs of entities —such as authors, keywords, or countries— appear together in the same document. Each entity corresponds to a node in the resulting graph, and edge weights reflect the strength of association between pairs. However, raw co-occurrence counts tend to be skewed by marginal frequencies: commonly occurring entities will naturally appear more often with others, even if no meaningful association exists. For this reason, a normalization step is essential to produce interpretable and visually coherent network representations.

Let $\mathcal{V} \equiv \{v_1, \ldots, v_N\}$ denote the set of entities and $M = [m_{ij}]$ the symmetric co-occurrence matrix, where m_{ij} is the total co-occurrence weight assigned to the pair (v_i, v_j) across all documents. For convenience, the diagonal stores the per-entity document frequencies f_i (number of documents in which v_i appears; full counting), so $m_{ii} = f_i$.

We adopt fractional counting for pairs: if a document contains p distinct entities, each unordered pair (v_i, v_j) with $i \neq j$ contributes $\omega^{\text{frac}}_{ij,(k)} = 1/\binom{p}{2}$, so the document's total pairwise contribution sums to 1. Aggregating over documents yields $m_{ij} = \sum_k \omega^{\text{frac}}_{ij,(k)}$ for $i \neq j$, while f_i is computed by full counting (one per document containing v_i). This mixed scheme (fractional for pairs, full for marginals) downweights very long lists within individual documents while preserving the interpretable binary notion of presence at the marginal level; a sensitivity check using an all-fractional variant for f_i produced qualitatively similar maps.

To mitigate the influence of high-frequency nodes and emphasize genuine associations, we set $m_{ii} = f_i$ for bookkeeping but ignore self-co-occurrences in subsequent analyses; the weighted adjacency is $W = (w_{ij})$ with $w_{ij} = m_{ij}/(f_i f_j)$, for $i \neq j$ and $w_{ii} = 0$. Alternative normalizations (for example, Salton/cosine, Jaccard, inclusion, equivalence) were inspected for sensitivity and led to qualitatively similar structures. Association strength is recommended for sparse scientific maps [31].

Once the weighted matrix W is computed, we remove the weakest links using a light threshold $\varepsilon = 0.05$ on w_{ij}. In this corpus, that eliminates roughly 10% of edges; sensitivity checks with $\varepsilon \in [0.03, 0.07]$ yielded qualitatively similar patterns. For keyword networks, we retain only the largest connected component to stabilize layouts. For the country-level co-authorship network, singleton nodes were retained to capture emerging or isolated collaborations; consequently, some centrality measures (such as closeness) are undefined for those nodes.

All steps were implemented in `R` with `bibliometrix` (data extraction), `igraph` (graph objects and metrics), and `ggraph` (visualization). A fixed random seed was set prior to layout optimization and community detection to ensure reproducibility. The final graph $G = (V, E, W)$ then feeds into the layout and community-detection steps described next.

3.3 *Network layout and community detection*

To embed the normalized graph in two dimensions we use the Kamada–Kawai spring algorithm [32]. For node v_i, let $\boldsymbol{x}_i = (x_i, y_i)^{\mathsf{T}}$ be its unknown coordinates. Then, the layout minimizes the expression given by

$$E = \frac{1}{2} \sum_{1 \leq i < j \leq N} k_{ij} \big(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \ell_{ij} \big)^2, \tag{11}$$

where we adopt a weighted variant in which the preferred distance is set to $\ell_{ij} = \ell_0 / w_{ij}$ (stronger ties imply shorter target distances) and the spring stiffness to $k_{ij} = \kappa / \ell_{ij}^2$. The sum stated in (11) includes only pairs with $w_{ij} > 0$, as zero–weight edges are removed during preprocessing.

Following [32], we set $\ell_0 = 1$ and $\kappa = 1$ after rescaling all edge weights w_{ij} to lie in $[0, 1]$. Other choices of ℓ_0 and κ, and an alternative construction with ℓ_{ij} proportional to weighted shortest–path distances, yielded visually similar layouts and did not affect community detection.

Communities are detected with the Louvain algorithm (modularity optimization, default resolution) [33], using a fixed random seed. Sensitivity checks with Walktrap produced qualitatively similar partitions.

Newton–Raphson iterations continue until the expression given in (11) stabilizes, placing strongly connected nodes close to one another.

For centrality metrics, betweenness is computed on the unweighted simple graph. Closeness uses weighted shortest–path distances $d_{ij} = 1/w_{ij}$ and is reported only for nodes in the largest connected component (nodes outside that component, including isolates, are not reported). PageRank is computed on the weighted, row–normalized adjacency matrix with the default damping factor 0.85 [34].

3.4 *Thematic mapping via centrality–density analysis*

Beyond a plain network plot, we summarize keyword relationships in a thematic map (strategic diagram) [35, 36]. After partitioning the keyword graph into clusters $T_1, \ldots, T_m$, each cluster is characterized by the Callon density and centrality stated as

$$D_\alpha = 100 \frac{\sum_{\substack{i<j \\ v_i, v_j \in T_\alpha}} w_{ij}}{\binom{|T_\alpha|}{2}}, \qquad C_\alpha = 10 \frac{\sum_{v_i \in T_\alpha} \sum_{v_j \notin T_\alpha} w_{ij}}{|T_\alpha|}, \tag{12}$$

which measure, respectively, internal cohesion and external connectedness. Singleton clusters are excluded, because D_α is undefined for $|T_\alpha| = 1$. Plotting the points (C_α, D_α) yields four quadrants: motor, niche, basic or transversal, and emerging or declining themes. Quadrants in the strategic diagram are defined by the median values of C_α and D_α computed across retained clusters.

This method —matrix construction, normalization, Kamada–Kawai layout, community detection, and strategic mapping— implemented with `bibliometrix`, `igraph`, and `ggraph`, allows us to explore large bibliographic corpora, identify influential thematic hubs, and track shifts in research priorities with quantitative rigor and visual clarity.

4. Empirical Results

This section presents the main findings from the bibliometric analysis, focusing on publication trends, journal and document types, and international collaboration patterns. The results reveal a research field characterized by steady but niche growth, a diverse journal landscape, and a globally distributed, yet loosely connected, network of contributors.

4.1 *Descriptive profile of the corpus*

The final corpus comprises 48 peer-reviewed articles published between 1999 and 2025. As summarized in Table 5, and in Figure 3, annual output remains modest (never exceeding four papers/year), with local peaks in 2015 and 2019 (four each). The apparent decline in 2025 should be read cautiously because our search window closes on 11 June 2025 (partial year). This temporal pattern is consistent with a technically specialized field where methodological contributions accumulate steadily but at low volume.

In terms of document type, 44 of the included articles are original research papers, three are review articles, and one is a conference proceeding. The predominance of original research reflects an active and ongoing interest in methodological development. Conversely, the small number of reviews suggests that synthesis-oriented contributions remain relatively limited, possibly due to the technical complexity and evolving scope of the field. This underscores the potential value of structured overviews, such as the present one, that aim to consolidate dispersed efforts across subdomains.

Table 5.: Number of publications per year (1999–2025, $n = 48$)

Year	# publications	Year	# publications	Year	# publications
1999	1	2012	1	2021	2
2000	1	2013	1	2022	2
2002	3	2014	2	2023	2
2003	2	2015	4	2024	1
2004	2	2016	1	2025	1
2006	1	2017	3		
2007	3	2018	3		

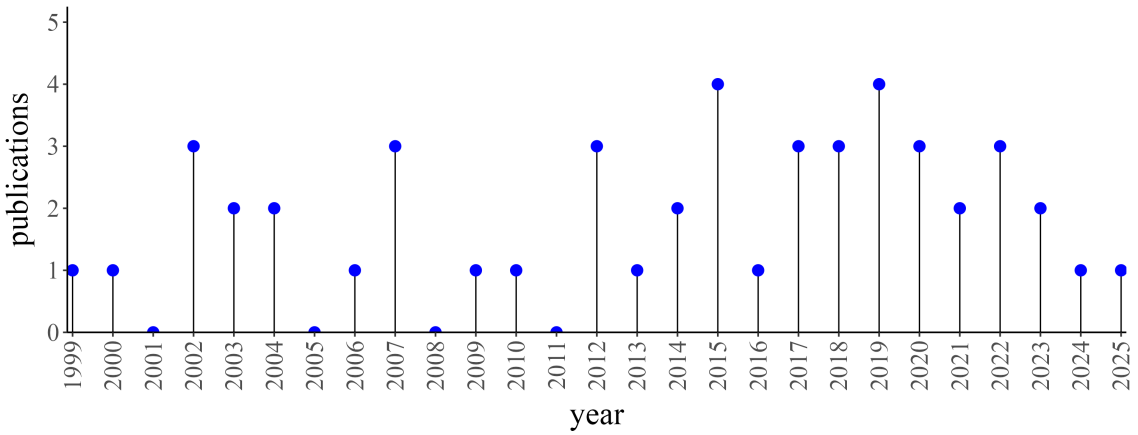

Figure 3.: Time-series plot of publications for 1999–2025.

The articles are distributed across 21 different journals, indicating that research on non-Markov multi-state models has found outlets in a range of applied statistics and biostatistics venues. Table 6 lists the eight journals that published at least two papers from the corpus. They account for 35 publications, or approximately 73% of the total. The three most frequent journals—Statistics in Medicine, Statistical Methods in Medical Research, and Lifetime Data Analysis—together represent exactly half of the corpus, highlighting their central role in disseminating methodological advances in this area. The remaining articles are spread across 13 other journals, reflecting a thematically diverse and methodologically distributed landscape.

Table 6.: Journals with at least two publications ($n = 48$).

Journal	# publications	Share
Statistics in Medicine	9	0.188
Statistical Methods in Medical Research	8	0.167
Lifetime Data Analysis	7	0.146
Biometrics	3	0.063
Biometrika	2	0.042
Communications in Statistics – Simulation and Computation	2	0.042
International Journal of Biostatistics	2	0.042
Scandinavian Journal of Statistics	2	0.042

Beyond yearly counts and outlets, it is useful to visualize how the primary orientation of contributions evolved over time. Table 7 presents a compact, snapshot by year, distinguishing methodological/seminal versus applied/case-driven articles. Note: Counts reflect the primary orientation of each paper and exclude review-only and proceedings items. Per-year counts have been recomputed to ensure they never exceed the annual totals in Table 5. Minor discrepancies across tables can occur due to deduplication and final metadata harmonization. To contextualize these counts with concrete methodological milestones and applications, Table 8 summarizes key advances across the same period.

Table 7.: Primary orientation by year: methodological/seminal versus applied/case-driven.

Year	Methodological/seminal	Applied/case-driven
1999	1	0
2000	1	0
2002	3	0
2003	1	0
2004	2	0
2006	1	0
2007	3	0
2009	1	0
2010	1	0
2012	2	1
2013	0	0
2014	2	0
2015	4	0
2016	1	0
2017	3	0
2018	3	0
2019	3	1
2020	2	1
2021	2	0
2022	1	0
2023	2	0
2024	1	0
2025	1	0

Table 8.: Timeline of key methodological and applied advances

Year(s)	Milestone
1999–2002	Nonparametric non-Markov transition/probability estimation [37]; IPCW for dependent censoring in multi-state systems [24]; flow-graph and saddlepoint machinery [38]; likelihood for interval-censored multi-state [23].
2006–2007	Consistent non-Markov estimators for illness–death [20]; landmarking for dynamic prediction [39]; software enablers (`TDC.MSM`) [40].
2010–2014	Semi-Markov with interval-censoring and GOF tools [41]; semiparametric multi-state for trial extrapolation [42].

(continued on the next page)

Table 8 (continued from previous page)

Year(s)	Milestone
2015	Robust non-Markov estimators and role of frailty [13, 11, 43].
2018	LAJ formalized with asymptotics [12].
2019	IPD-free empirical-hazard bootstrap [44]; comprehensive software for semi-competing risks [45]; cost-effectiveness comparison (applied) [46].
2020	IPCW product-limit estimators and local tests for non-Markov settings [47]; dynamic prediction in practice [48].
2022	Unified review of non-Markov inference and pseudo-observation regression [14].
2023–2025	Scalability for massive EHR via optimal subsampling [16]; tensor-based landmarking [27]; simulation guidance for semi-Markov state occupancy [49].

To gain further insight into the structure of the research community itself, we now turn to patterns of international collaboration.

4.2 *Country collaboration network*

To explore patterns of international collaboration in this field, we constructed a country-level co-authorship network based on the 48 articles in the corpus. Each article was assigned to all countries represented in the authors' institutional affiliations, and an undirected edge was created between each pair of countries co-authoring at least one publication.

Edge weights are given by the association-strength index derived from raw co-authorship counts (Section 3.2), which highlights unexpectedly strong bilateral ties after accounting for marginal prolificness. The network is projected onto the plane using the Kamada–Kawai layout, and communities are identified via the Walktrap algorithm [30].

Table 9 summarizes the composition of each detected community and reports three standard centrality measures: betweenness (computed on the unweighted graph), and closeness and PageRank (based on the weighted graph). For countries that are not part of the main connected component, closeness centrality is undefined and recorded as `NaN`.

To provide a more granular view of research output and collaboration patterns, Table 10 reports additional metrics by country. For each nation, the total number of publications is disaggregated into single-country papers (SCP) and multi-country papers (MCP). The proportion of collaborative output is summarized by the ratio $\mathrm{MCP\_ratio} = \mathrm{MCP}/\mathrm{articles}$, which quantifies the degree of international co-authorship. By construction, $\mathrm{SCP} + \mathrm{MCP} = \mathrm{articles}$ and $\mathrm{MCP\_ratio} \in [0, 1]$, where 0 indicates no international co-authorship and 1 indicates exclusively international co-authorship.

In addition to these tabulated indicators, Figure 4 visualizes the global structure of collaboration. In Figure 4, node size encodes the number of articles affiliated with each country and colors indicate Walktrap communities. The clusters reflect within-community ties after association-strength normalization. The United States, France, and the Netherlands emerge as prominent contributors in terms of both publication volume and network connectivity. Countries such as Portugal, Spain, the United Kingdom, Australia, and (to a lesser extent) Germany exhibit a balanced profile, combining single-country publications with international co-authorship. In contrast, Brazil, Chile, Denmark, Iran, New Zealand, Slovenia, and Sweden rely entirely on international co-authorship ($\mathrm{MCP\_ratio} = 1$), whereas China, Finland, India, and Norway appear only with single-country publications ($\mathrm{MCP\_ratio} = 0$), forming peripheral nodes without active international links in this corpus.

Isolated nodes and small peripheral clusters suggest emerging engagement or structural barriers to deeper integration. In this context, international collaboration appears to play a vital role in sustaining methodological development across regions, particularly for smaller or less prolific research communities.

Table 9.: Community assignment and centrality statistics.

Country	Cluster	Betweenness	Closeness	PageRank
France	1	4	0.200	0.0908
United Kingdom	1	3	0.167	0.0681
Australia	1	0	0.143	0.0615
New Zealand	1	0	0.143	0.0615
Sweden	1	0	0.111	0.0386
Netherlands	2	2	0.250	0.0832
Spain	2	2	0.250	0.0832
Portugal	2	0	0.167	0.0450
Germany	2	0	0.167	0.0450
USA	3	3	0.333	0.1230
Iran	3	0	0.200	0.0445
Brazil	3	0	0.200	0.0445
Chile	3	0	0.200	0.0445
Denmark	4	0	-	0.0641
Slovenia	4	0	-	0.0641
China	5	0	-	0.0096
Norway	6	0	-	0.0096
Finland	7	0	-	0.0096
India	8	0	-	0.0096

Closeness is computed on the largest connected component using weighted shortest paths ($d_{ij} = 1/w_{ij}$); entries marked '-' are undefined for isolates or small disconnected components.

Additional country-level details are summarized in Table 10. Overall, the development of non-Markov multi-state survival modeling is sustained by a globally distributed but structurally sparse set of contributors. While a few countries account for a substantial share of the literature, selective international partnerships extend the reach of the field and enable participation from emerging centersof expertise.

4.3 *Keyword analysis and thematic map*

To complement the structural analysis of contributors, we examined the thematic composition of the field through author-supplied keywords. Synonyms were harmonized using standard stemming and lowercasing procedures to facilitate aggregation. The frequency distribution is summarized in Figure 5, which displays the twenty most common keywords across the 48-article corpus. Foundational terms (multi-state model, survival analysis) dominate, with substantial presence of semi-Markov model and landmark analysis, indicating the dual emphasis on relaxing Markov assumptions and enabling dynamic prediction.

A complementary word cloud is presented in Figure 6, offering an at-a-glance view of keyword salience: larger fonts mark more frequent author-supplied terms. While it does not encode co-occurrence structure (which is captured by the network and thematic map), it complements Figure 5 by giving an intuitive sense of the vocabulary used across the corpus. The keyword analysis reveals the following consistent thematic strands:

- Core modeling frameworks —Terms such as multi-state model, semi-Markov model, and survival analysis indicate that foundational statistical structures remain central to the field's identity.
- Estimation methods —Frequent mention of estimators such as AJ, Kaplan–Meier, and Cox proportional hazards suggests continued reliance on traditional tools, often adapted to the non-Markov setting.

Table 10.: Publication and collaboration metrics by country.

Country	Publications	SCP	MCP	MCP_ratio	Relative frequency
USA	14	11	3	0.21	0.29
Netherlands	8	6	2	0.25	0.17
France	7	5	2	0.29	0.15
Spain	6	3	3	0.50	0.13
Portugal	4	2	2	0.50	0.08
United Kingdom	4	2	2	0.50	0.08
Germany	3	2	1	0.33	0.06
Australia	2	1	1	0.50	0.04
Brazil	1	0	1	1.00	0.02
Chile	1	0	1	1.00	0.02
China	1	1	0	0.00	0.02
Denmark	1	0	1	1.00	0.02
Finland	1	1	0	0.00	0.02
India	1	1	0	0.00	0.02
Iran	1	0	1	1.00	0.02
New Zealand	1	0	1	1.00	0.02
Norway	1	1	0	0.00	0.02
Slovenia	1	0	1	1.00	0.02
Sweden	1	0	1	1.00	0.02

where, as mentioned, SCP is single-country publications and MCP is multi-country publications, where the relative frequency is the share of publications affiliated with a given country, with the column sums exceeding the value of one because multi-authored papers are counted for every participating country.

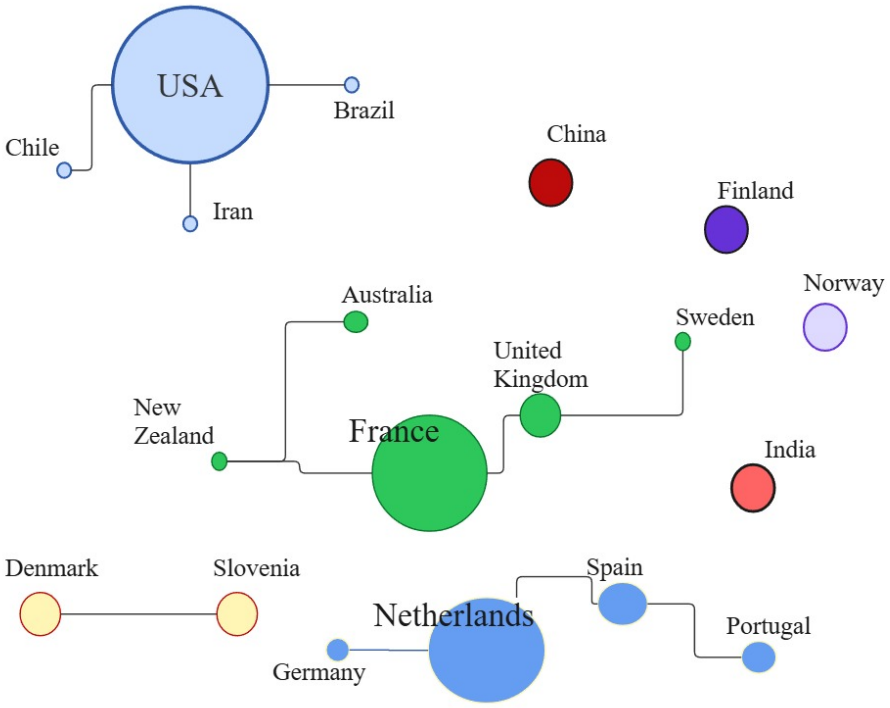

Figure 4.: Country collaboration network, where the node size is proportional to article count and colors indicate Walktrap communities.

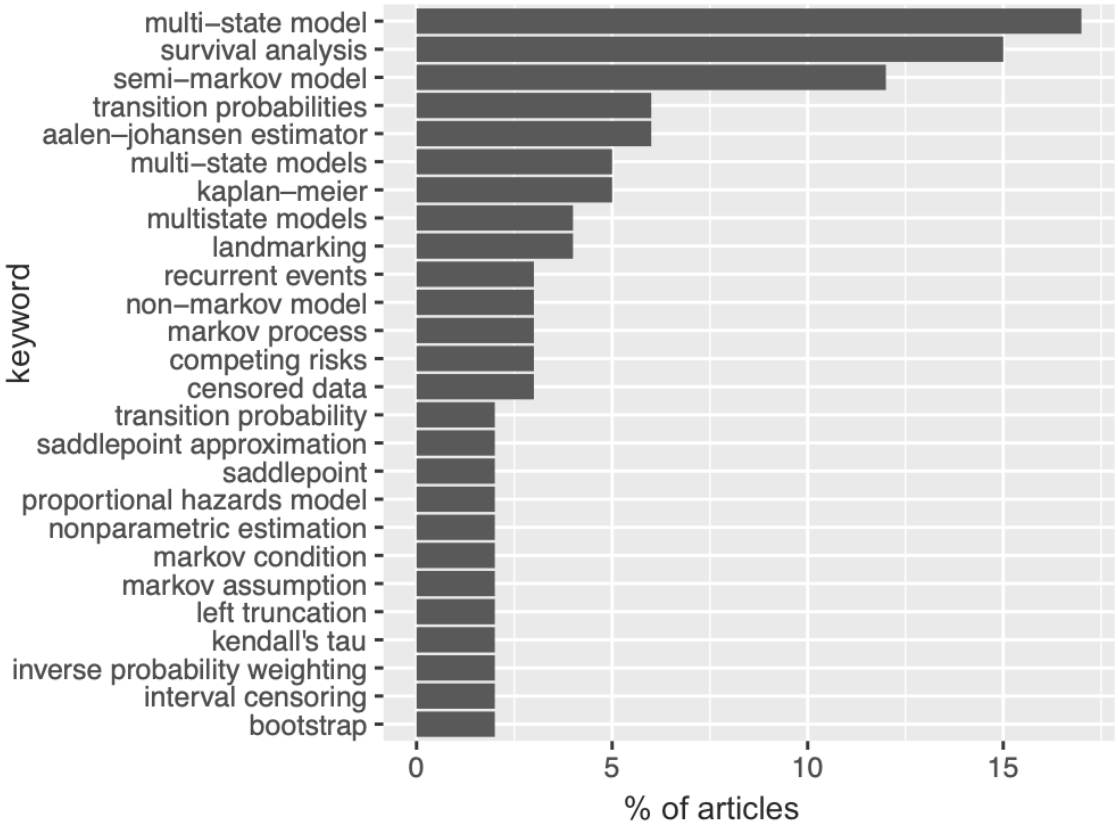

Figure 5.: Twenty most frequent keywords (in percentage of articles in which the keyword appears).

- Predictive modeling —Keywords like landmark analysis, dynamic prediction, and time-varying covariates reflect increasing interest in personalized prognostics and temporally adaptive estimation.
- Computational techniques —Entries such as saddlepoint approximation, bootstrap, and subsampling point to an emphasis on scalability and numerical efficiency, particularly for large or irregular datasets.

The keyword analysis suggests a research field that combines stable methodological foundations with focused innovations. The prevalence of terms related to prediction, flexibility, and computation points to a gradual diversification of priorities, potentially indicating a shift toward more application-aware and data-intensive methodologies. Whether these developments will ultimately converge into a unified modeling framework or continue to evolve in parallel remains an open question for future investigation.

To explore how these thematic strands relate to one another, we constructed a strategic diagram based on the co-occurrence network of author-supplied keywords (see Section 3.4). Using the Louvain algorithm for community detection, we identified keyword clusters within the network. The results were consistent with those obtained via the Walktrap algorithm, and only clusters containing at least three keywords were retained.

Figure 6.: Word cloud of the most frequent author keywords.

For each retained cluster T_α, we computed Callon centrality C_α and Callon density D_α, as defined in (12). These metrics quantify, respectively, the degree of external connectivity of a cluster to the rest of the network and its internal cohesion. Seven clusters met the inclusion criteria, and their corresponding centrality and density scores are reported in Table 11. By construction, $D_\alpha \in [0, 100]$. Centrality $C_\alpha \geq 0$; with association-strength weights and our sparsification, observed C_α values lie well below 10 in this corpus.

Table 11.: Keyword clusters on the strategic diagram ($n = 48$ papers).

Cluster	Most representative keyword	C_α	D_α
1	AJ estimator	0.77	49.1
2	Multi-state model	1.74	13.8
3	Cox proportional hazards model	0.81	54.9
4	Illness–death model	0.31	25.0
5	Landmark analysis	0.59	47.6
6	Saddlepoint approximation	0.13	82.5
7	Left truncation	0.00	50.0

The resulting clusters are visualized in Figure 7, where each point represents a keyword cluster positioned by Callon centrality (C, external connectedness) and density (D, internal cohesion); dashed lines indicate median C and D. Clusters for the Cox/proportional hazards model and the AJ estimator lie in the motor quadrant (high C, high D), confirming their role as methodological anchors. Topics such as saddlepoint approximation appear as niche (high D, low C), reflecting strong internal development but limited cross-theme ties.

The multi-state model cluster is basic/transversal (higher C, lower D), acting as a broad umbrella rather than a tightly knit topic. Themes like illness–death plot as emerging/declining (low C, low D), signaling either a shift in emphasis or early-stage consolidation.

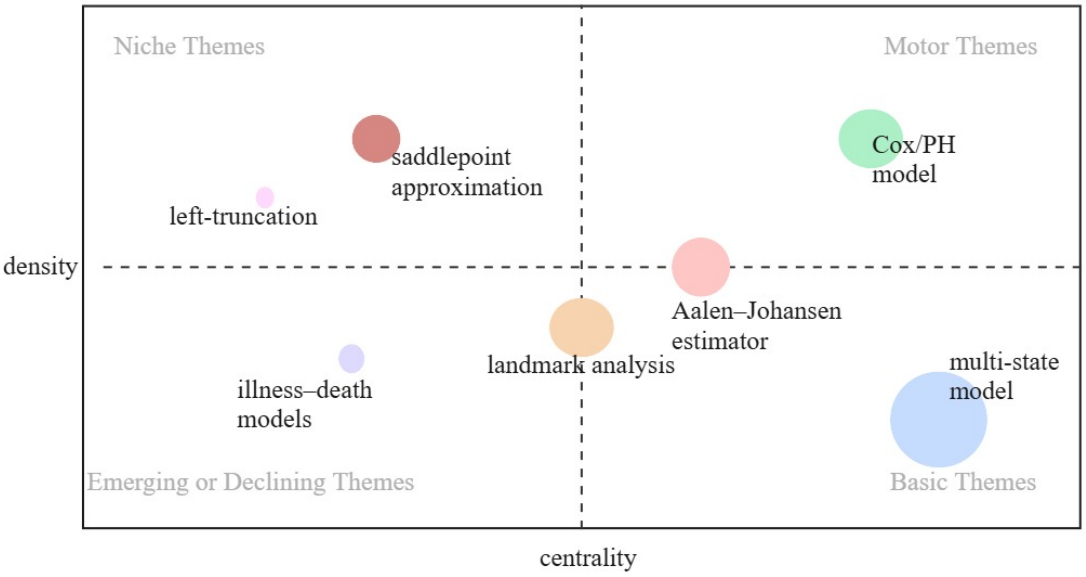

Figure 7.: Strategic diagram of keyword clusters. Axes correspond to Callon centrality (C) and density (D); dashed lines indicate the median values.

When compared with the raw keyword frequencies in Figure 5, the strategic map offers complementary insights. For instance, while multi-state model is the most frequently occurring keyword, its low density score highlights its function as a broad, umbrella term rather than a tightly defined topic. In contrast, clusters focused on Cox/proportional hazards and AJ estimators, though less frequent, exhibit high internal consistency and strong connectivity to other themes, underscoring their methodological maturity. Other developments —such as saddlepoint approximation— remain peripheral in terms of network centrality but show high internal cohesion, suggesting sustained attention within specialized communities. The central positioning of landmark analysis and time-varying covariates further aligns with the increasing interest in dynamic prediction frameworks.

Overall, the thematic map reflects a field anchored in traditional survival methods while gradually incorporating more flexible, data-adaptive, and computationally intensive approaches. Whether peripheral topics such as left truncation or high-resolution inference will eventually integrate into a unified modeling framework or continue to evolve as specialized subfields remains an open avenue for future exploration.

5. Discussion

This section reflects on the methodological contributions and bibliometric patterns observed in the reviewed literature. We summarize the thematic evolution of modeling approaches and highlight priorities for future development and practical implementation.

5.1 *Methodological and conceptual synthesis*

Recent related developments in flexible survival modeling also include distributional innovations designed to accommodate a wide range of hazard shapes and censoring mechanisms. For example, [73] introduced an exponentiated–Weibull–G family with regression and simulation tools for censored COVID-19 data, illustrating parallel goals of flexibility and applied validation within survival analysis.

The body of literature analyzed illustrates a steady progression from traditional survival models to more flexible, application-aware frameworks. Many contributions extend foundational approaches —such as the Cox model and the AJ estimator— by relaxing key assumptions to better accommodate time-dependent dynamics, path dependence, and complex censoring structures. Notable generalizations include semi-Markov transition intensities [50, 51], non-Markov illness–death and semi-competing risks structures [20, 11], and models incorporating subject-level frailties [43], especially relevant for joint modeling of progression-free and overall survival.

Across studies, estimation strategies tend to favour nonparametric and semiparametric methods, valued for their robustness and interpretability in the presence of censoring irregularities. Common approaches include weighted NA estimators [24], penalized likelihood methods [23], pseudo-observation frameworks [14], and Bayesian MCMC schemes [45]. Many of these techniques are designed to align with estimand definitions under the ICH E9 (R1) [4, 5], particularly in the context of late-phase clinical trials or real-world registry-based studies.

From a computational standpoint, progress has been incremental rather than transformative. Advances such as saddlepoint approximations [38, 42], rare-event subsampling techniques [16], and bootstrap-based procedures for hazard estimation [44] have improved efficiency, yet full likelihood-based modeling remains computationally challenging in large-scale EHR cohorts. The development of dedicated software —such as `TDC.MSM` [40], `msm` [7], and `SemiCompRisks` [45]— has played an important role in dissemination, reproducibility, and methodological uptake.

The applied motivations behind these methodological innovations are diverse. Case studies in oncology [13, 52], organ transplantation [41, 53], infectious diseases [54], and population-scale EHR data [16] have revealed limitations in survival methods. These applications have prompted the adoption of models capable of accommodating intermediate events, irregular observation schemes, and dynamic prediction targets. The overall picture is one of methodological diversification, oriented toward frameworks that are statistically coherent, computationally feasible, and aligned with the structure of real-world data.

Nevertheless, several areas remain underdeveloped. Few methods have been benchmarked systematically across diverse clinical or data environments, and the lack of shared test beds hinders comparative evaluation. Models incorporating high-dimensional covariates—particularly those from omics platforms, medical imaging, or wearable devices—are still rare despite growing interest. Although tools for assessing the validity of the Markov and semi-Markov assumptions exist [47, 11, 14], they are not yet standardized or routinely applied in practice. Expanded simulation-based studies [49], harmonized modeling workflows, and accessible diagnostic procedures would enhance transparency and comparability, particularly in regulatory or decision-analytic settings.

5.2 *Structural patterns and practical implications*

The bibliometric structure of the field reveals a fragmented yet internationally distributed research landscape. As discussed in Section 4.2, collaboration networks are strongest among countries such as the United States, France, and the Netherlands. Nonetheless, smaller but meaningful contributions from nations including Portugal, Iran, and Brazil reflect a growing global engagement. The dispersion of publication venues and the co-occurrence of keywords further underscore a thematically diffuse field, marked by methodological specialization but limited integration across strands.

The strategic diagram of thematic clusters reinforces this view, placing multi-state models at the conceptual centerof the field while highlighting their heterogeneous interpretations. Canonical methods —such as the Cox model and the AJ estimator— remain foundational, yet extensions involving semi-competing risks, illness–death structures, and landmark prediction frameworks [12] are gaining prominence. Other computational innovations —including tensor-based dynamic prediction [27] and saddlepoint approximations [38]— show technical promise, but their integration into standard workflows is still limited.

On the practical side, the proliferation of `R` packages —such as `multi-state` [55], `msm` [7], and `SemiCompRisks` [45]— has greatly improved access to advanced models. However, software interoperability remains a challenge. Differing assumptions, input specifications, and diagnostic capabilities across packages can produce inconsistent results, particularly when applied without close statistical oversight. A harmonized interface, clearer documentation of modeling assumptions, and shared test scenarios would help to reduce analytical variability and promote reproducibility.

For a practitioner-oriented overview of available `R` tools aligned with the methods discussed in Sections 2 and 3, Table 12 summarizes the scope, key functionality, assumptions, and practical notes of widely used packages.

Table 12.: Software landscape for (non-)Markov multi-state analysis: scope, key functionality, assumptions, and practical notes.

`R` package	Scope and models	Key functionality	Typical assumptions/limitations	Scalability and practical notes
`mstate`	General multi-state and competing risks (Markov baseline) with strong data-prep tools.	`msprep` (wide→long), risk sets, AJ via `probtrans`, Cox-based semi-parametric fits, plotting/vignettes.	Non-Markov analyses typically require external strategies (such as, landmarking, plug-in).	Stable baseline ecosystem; excellent pedagogy; widely used in clinical examples.
`msm`	Continuous-time multi-state models for panel data (Markov baseline).	Transition-intensity modeling; covariates; prediction; fitted-model summaries.	Primarily Markov workflows; non-Markov behavior needs external strategy.	Mature, widely used; good baseline toolkit.
`etm`	Nonparametric estimation of transition probabilities (AJ/etm) and related quantities.	AJ, state-occupation probabilities, variance estimation.	Focus on panel observations; non-Markov requires landmark/plug-in outside core functions.	Lean implementation; good for teaching/benchmarking AJ.
`TDC.MSM`	Multi-state with time-dependent covariates; supports Markov, semi-Markov, piecewise, and nonparametric variants.	Transition probabilities, hazards, survival; plotting utilities.	Mainly right-censoring; users must align data structure and model choice.	Practical for applied analyses; often cited in non-Markov discussions.

Continued on the next page

Table 12 (continued from previous page)

R package	Scope and models	Key functionality	Typical assumptions/limitations	Scalability and practical notes
multi-state	Extensions for illness–death/ interval-censored settings; IPW workflows.	Transition probabilities and IPW methods for interval-censored data.	Validity hinges on correct censoring/weight models; targeted designs.	Useful when interval censoring + confounding are present.
TPmsm	Transition probabilities for 3-state progressive models (incl. non-Markov).	Multiple nonparametric/semiparametric estimators; IPCW options; worked examples.	Geared to 3-state structures; careful landmark/windowing and censoring models needed.	Strong when the target is $P(s,t)$ under history dependence.
survIDM	Illness–death under Markov (clock-forward) and semi-Markov (clock-reset).	Transition and state-occupation probabilities, cause-specific hazards, prediction utilities.	Focus on illness–death graphs; other topologies limited; assumptions must match design.	Balanced inference and prediction for illness–death; detailed documentation.
SemiCompRisks	Semi-competing risks/illness–death (independent or clustered); Markov/semi-Markov.	Proportional hazards/accelerated failure time submodels; parametric and nonparametric baselines; random effects; Bayesian MCMC (and some maximum likelihood estimates).	Mainly right-censoring; MCMC cost can be non-trivial in large cohorts.	Versatile for clinical semi-competing risks; simulation utilities included.

5.3 *Practical scalability, outlook, and answers to the research questions*

Beyond methodological considerations, computational scalability and software readiness remain binding constraints. Subsampling and parallelization strategies [16] can mitigate wall-time, but full-likelihood implementations still scale poorly when left truncation, interval censoring, and large, irregular EHR cohorts interact. Pragmatic routes forward include graphics processing unit (GPU)-enabled Bayesian samplers and distributed or composite-likelihood schemes, where likelihoods are heavy, saddle-point or surrogate-objective approximations can deliver credible uncertainty at a fraction of the cost. Just as crucial are standardized data-preparation and risk-set constructors (Table 12), together with transparent IPCW methods, because these operational pieces often dominate bias–variance trade-offs as much as the nominal estimator itself.

To close the loop with the objectives articulated in the section of introduction, we now revisit RQ1–RQ5 and summarize how the review, the mini-simulation, and the bibliometric analysis jointly answer them.

RQ1 (methods/estimands) When the Markov assumption is doubtful, practice converges on three families: (i) LAJ; (ii) non-Markov plug-in estimators for $P_{rj}(s,t)$ based on conditional cumulative incidence; and (iii) semi-Markov intensity models capturing sojourn-time effects. Under history-dependent censoring, IPCW is essential regardless of the estimator (Section 2; Table 3).

RQ2 (finite-sample performance) In our semi-Markov simulation with history-dependent censoring, both baseline and LAJ overestimated $P_{rj}(s,t)$; landmarking reduced bias but did not eliminate it in the absence of IPCW, and later landmarks increased RMSE due to smaller risk sets (Section 2.4; Table 4).

RQ3 (software) Widely used packages (`mstate`, `msm`, `survIDM`, `TPmsm`, `SemiCompRisks`, and others) cover Markov baselines well and provide routes to non-Markov analyses via landmarking, plug-ins, or model-based approaches, but interoperability and explicit documentation of modeling assumptions remain uneven (Table 12).

RQ4 (field structure) The bibliometric network shows a globally distributed yet fragmented field with methodological anchors around Cox/proportional hazards and AJ; niche but cohesive topics (such as saddlepoint approximations) remain peripheral; collaboration hubs include the USA, France, and the Netherlands (Section 4; Figures 4, 7).

RQ5 (implementation priorities) Priorities include routine diagnostics for history dependence (AJ versus landmark contrasts; sojourn-stratified hazards), principled handling of censoring via IPCW, harmonized workflows aligned with the ICH E9(R1) estimand framework, and scalable, auditable software for large EHR cohorts (Sections 2 and 5).

Overall, the literature portrays a methodologically rich yet structurally fragmented field. Non-Markov models now support applications ranging from clinical-trial design to health-technology assessment and personalized prognosis. Broad uptake, however, hinges on three enablers: (i) computational efficiency at EHR scale; (ii) robust diagnostics aligned with the ICH E9(R1) estimand framework; and (iii) auditable, interoperable software and workflows (Tables 3 and 12). The decision rules, estimator comparisons, and software map provided here are intended as practical scaffolding for applied teams.

This synthesis is restricted to English-language, peer-reviewed articles indexed in Web of Science and Scopus and relies on bibliometric metadata; grey literature may therefore be missed and some affiliations misclassified. The mini-simulation prioritizes pedagogy over breadth (single progressive graph, no IPCW), illustrating mechanisms rather than benchmarking performance. We mitigate these limits by releasing the analysis code and the exact database queries (Appendices A and B), enabling replication and extension.

6. Conclusions

Non-Markov multi-state survival analysis has evolved from a primarily theoretical construct into a practical extension of conventional time-to-event models. Its capacity to represent recurrent events, intermediate transitions, and complex censoring structures has increased its relevance in contemporary clinical and epidemiological research. This review, supported by a complementary bibliometric analysis, has outlined both the methodological depth and the growing —though uneven— adoption of these models across domains such as oncology, transplantation, and real-world studies based on electronic health records.

Our key take-aways are the following:

(i) Non-Markov targets need conditional machinery —Baseline AJ is a robust benchmark for Markov-like settings, but under history dependence one should use LAJ or conditional plug-in estimators (with the increments computed in the landmark subcohort).

(ii) Censoring must be modeled when history-dependent —Both baseline and landmark estimators remain biased if dropout depends on past history; IPCW (with diagnostics) is needed for valid inference.

(iii) Diagnostics before commitment —Contrast AJ versus landmark estimates, stratify hazards by sojourn time, and check sensitivity to the visit process and weights; align estimands with ICH E9(R1) and ensure risk-set construction respects delayed entry.

(iv) Software exists (workflows matter) —The software landscape covers most use-cases, but interoperability and transparent assumptions are decisive; shared test beds and auditable methods reduce analysis variability.

Looking forward, several areas stand out as priorities for further refinement. First, the development and standardization of diagnostic tools for assessing departures from the Markov assumption remain essential. More consistent use of graphical techniques, hypothesis tests, and resampling-based procedures would strengthen the justification of modeling choices in applied contexts.

Second, scalability remains a limiting factor in many real-world applications. The use of parallel computing, rare-event subsampling, and GPU-enabled Bayesian inference may help to address computational bottlenecks, particularly in the analysis of large or high-dimensional data.

Third, closer integration with machine learning approaches could support the application of non-Markov methods to data derived from omics, imaging, or real-time monitoring. This may involve coupling survival models with techniques for feature selection, representation learning, and uncertainty quantification. Finally, progress toward consensus validation —through open benchmarks, shared datasets, and reproducible workflows— would help to harmonize the current diversity of methods and support evidence-based methodological guidance, as has occurred in related areas such as joint modeling and competing risks.

Collectively, advances in these areas may contribute to a more unified, interpretable, and accessible framework for non-Markov survival modeling. Such progress would support broader methodological uptake in trial design, regulatory science, personalized risk prediction, and health-technology assessment. While further consolidation is needed, the field already provides a strong foundation for addressing the temporal and structural complexity of modern biomedical data.

Appendix A. R code for the simulation

```
## 0) Packages -------------------------------------------------------------
req <- c("survival","mstate","ggplot2","dplyr","tidyr","purrr",
         "readr","xtable","gridExtra")
inst <- rownames(installed.packages())
if (length(setdiff(req, inst))>0) install.packages(setdiff(req, inst))
invisible(lapply(req, require, character.only = TRUE))

## 1) Output directory -----------------------------------------------------
outdir <- "sim_out"
if (!dir.exists(outdir)) dir.create(outdir, recursive = TRUE)

## 2) Global settings ------------------------------------------------------
set.seed(20250831)
n_per_rep   <- 500          # sample size per replicate
n_reps      <- 200          # MC replicates (use 5 for a quick test)
times_eval  <- c(6,12,18,24)  # evaluation times
s_landmark  <- c(6,12)        # landmark times
tau_admin   <- max(times_eval) # administrative end

## Transition and censoring parameters:
```

```
## 1->2 : 0.15 ; 1->3 : 0.05
## 2->3 : h23(w, T_P) = c*(w+1)^0.3, with c = 0.08*exp(0.4 * I(T_P>6))
## Censoring: lambda_C(t) = 0.02*exp(0.6 * I(t>T_P))
h12 <- 0.15
h13 <- 0.05
c0  <- 0.08
alpha_Tp <- 0.4
pow <- 0.3
hc0 <- 0.02
alpha_c <- 0.6

## 3) Sojourn sampler in state 2 (inverse-CDF) ------------------------------
## Draws time spent in state 2 (w), given c and power 'pow'.
r_sojourn <- function(n, cval, pow=0.3) {
  U <- runif(n)
  w <- (1 - (pow + 1)/cval * log(U))^(1/(pow + 1)) - 1
  pmax(w, 0)
}

## 4) Helpers: sanitization and tiny zero-length fixes ----------------------
## Ensure numeric times, integer statuses (0/1), and no non-finite values.
.ms_sanitize_df <- function(df, time_cols = NULL, status_cols = NULL) {
  if (is.null(time_cols)) {
    time_cols <- intersect(c("t12_abs","t13_abs","t23_abs","t2_abs","t3_abs"),
                           names(df))
  } else {
    time_cols <- intersect(time_cols, names(df))
  }
  if (is.null(status_cols)) {
    status_cols <- intersect(c("status12","status13","status23","status2",
                               "status3"), names(df))
  } else {
    status_cols <- intersect(status_cols, names(df))
  }
  for (nm in time_cols) {
    v <- suppressWarnings(as.numeric(df[[nm]]))
    v[!is.finite(v)] <- NA_real_
    df[[nm]] <- v
  }
  for (nm in status_cols) {
    v <- as.integer(df[[nm]])
    v[is.na(v)] <- 0L
    df[[nm]] <- v
  }
  df
}

## Rare corner case: force Tstop > Tstart for events at time 0.
.fix_zero_intervals <- function(msdata) {
  if (!all(c("Tstart","Tstop","status") %in% names(msdata))) return(msdata)
  bad <- which(msdata$status == 1 & msdata$Tstop <= msdata$Tstart)
  if (length(bad)) {
    eps <- 1e-8
    msdata$Tstop[bad] <- msdata$Tstart[bad] + eps
  }
  msdata
}

## 5) Cohort simulator (returns true times + censoring) ---------------------
## Keep the "true" times; build mstate inputs later.
simulate_cohort <- function(n, with_censor=TRUE) {
```

```
  ## True event times (no censoring) from state 1
  t12_true <- rexp(n, rate = h12)  # time to progression
  t13_true <- rexp(n, rate = h13)  # time to direct death
  prog_first <- (t12_true < t13_true)

  ## Death after progression (absolute time)
  cval <- c0 * exp(alpha_Tp * as.numeric(t12_true > 6))
  w23  <- r_sojourn(n, cval=cval, pow=pow)
  t23_true_abs <- ifelse(prog_first, t12_true + w23, NA_real_)

  ## History-dependent censoring + administrative cut
  if (with_censor) {
    C <- numeric(n)
    for (i in seq_len(n)) {
      Tp <- if (prog_first[i]) t12_true[i] else Inf
      c1 <- rexp(1, rate = hc0)
      if (c1 < Tp) {
        C[i] <- c1
      } else {
        if (is.finite(Tp)) {
          c2 <- rexp(1, rate = hc0 * exp(alpha_c))
          C[i] <- Tp + c2
        } else {
          C[i] <- c1
        }
      }
    }
  } else {
    C <- rep(Inf, n)
  }
  C <- pmin(C, tau_admin)

  data.frame(
    id = seq_len(n),
    t12_true = t12_true,
    t13_true = t13_true,
    t23_true_abs = t23_true_abs,
    cens_abs = C
  )
}

## 6) Build "state-wide" times/status for msprep() --------------------------
## Columns by STATE (not by transition):
## - t2_abs/status2 : time/indicator of entry into state 2 (progression)
## - t3_abs/status3 : time/indicator of entry into state 3 (death)
make_state_from_true <- function(wide_true) {
  ## Observed death time (direct or after progression) or censoring
  t3 <- pmin(
    wide_true$t13_true,
    ifelse(is.na(wide_true$t23_true_abs), Inf, wide_true$t23_true_abs),
    wide_true$cens_abs
  )
  status3 <- as.integer(t3 < wide_true$cens_abs)

  ## Observed time until first event leaving state 1
  t2 <- pmin(wide_true$t12_true, t3)
  status2 <-as.integer(wide_true$t12_true<=t3&
                         wide_true$t12_true<=wide_true$cens_abs)

  .ms_sanitize_df(data.frame(
    id       = wide_true$id,
```

```
    t2_abs   = t2,      status2 = status2,
    t3_abs   = t3,      status3 = status3
  ))
}

## 7) AJ from state-wide (generic) ----------------------------------------
aj_from_statewide <- function(state_df, times, predt = 0) {
  tmat <- mstate::transMat(x = list(c(2,3), c(3), c()),
                           names = c("1","2","3"))
  msdata <- mstate::msprep(
    data   = state_df,
    trans  = tmat,
    time   = c(NA, "t2_abs", "t3_abs"),    # times per STATE
    status = c(NA, "status2","status3"),  # status per STATE
    keep   = c("id")
  )
  ## Ensure Tstop > Tstart on event rows
  msdata <- .fix_zero_intervals(msdata)

  if (sum(msdata$status, na.rm = TRUE) == 0L) {
    return(dplyr::tibble(s = predt, t = times + predt, est = NA_real_))
  }

  cox <- survival::coxph(
    Surv(Tstart, Tstop, status) ~ strata(trans),
    data = msdata, ties = "breslow",
    model = FALSE, x = FALSE, y = FALSE,
    control = survival::coxph.control(timefix = FALSE)
  )

  msf <- mstate::msfit(cox, trans = tmat)
  pt  <- mstate::probtrans(msf, predt = predt)

  if (length(pt) == 0 || nrow(pt[[1]]) == 0) {
    ests <- rep(NA_real_, length(times))
  } else {
    ## Probability of being in state 3
    dfpt  <- pt[[1]]
    pcols <- grep("^pstate", names(dfpt), value = TRUE)
    if ("pstate3" %in% names(dfpt)) {
      p3 <- dfpt[["pstate3"]]
    } else if ("3" %in% names(dfpt)) {     # rare case
      p3 <- dfpt[["3"]]
    } else if (length(pcols)) {
      p3 <- dfpt[[tail(pcols, 1)]]
    } else {
      p3 <- rep(NA_real_, nrow(dfpt))
    }

    timegrid <- dfpt$time
    ests <- vapply(times, function(tt) {
      idx <- max(which(timegrid <= tt))
      if (!is.finite(idx) || idx < 1) NA_real_ else p3[idx]
    }, numeric(1))
  }
  dplyr::tibble(s = predt, t = times + predt, est = as.numeric(ests))
}

## 8) Baseline AJ (s = 0) -------------------------------------------------
aj_from0 <- function(wide_true, times = times_eval) {
  state_df <- make_state_from_true(wide_true)
```

```
  res <- aj_from_statewide(state_df, times = times, predt = 0)
  dplyr::mutate(res, method = "AJ_0")
}

## 9) Landmark AJ at time s ------------------------------------------------
landmark_aj <- function(wide_true, s, times = times_eval) {
  ## Subjects in state 1 and under observation at s
  in_s1 <- with(wide_true,
                ((is.na(t12_true) | t12_true > s) &
                   (is.na(t13_true) | t13_true > s) &
                   cens_abs > s))
  sub <- wide_true[in_s1, , drop = FALSE]
  if (nrow(sub) == 0) {
    return(dplyr::tibble(method = paste0("AJ_LM_s", s),
                         s = s, t = times[times > s], est = NA_real_))
  }

  ## Times relative to s
  t12a <- ifelse(!is.na(sub$t12_true)&sub$t12_true>s, sub$t12_true-s,Inf)
  t13a <- ifelse(!is.na(sub$t13_true)&sub$t13_true>s,sub$t13_true-s,Inf)
  t23a<-ifelse(!is.na(sub$t23_true_abs)&sub$t23_true_abs>s,
               sub$t23_true_abs-s,Inf)
  censa <- pmax(sub$cens_abs-s,0)

  ## State 3 after s
  t3a      <- pmin(t13a, t23a, censa)
  status3a <- as.integer(t3a < censa)

  ## State 2 after s
  t2a <- pmin(t12a, t3a)#if death before progression,time = death/censoring
  status2a <- as.integer(is.finite(t12a) & (t12a <= t3a) & (t12a <= censa))

  state_rel <- .ms_sanitize_df(data.frame(
    id       = sub$id,
    t2_abs   = t2a, status2 = status2a,
    t3_abs   = t3a, status3 = status3a
  ))

  ## AJ from new origin (0), then shift times back to calendar scale
  res <- aj_from_statewide(state_rel, times = times[times > s] - s, predt = 0)
  res$method <- paste0("AJ_LM_s", s)
  res$s <- s
  res$t <- res$t + s
  res
}

## 10) Monte Carlo truth (no censoring) ------------------------------------
## Returns method="TRUTH", s in {0, s_landmarks}, t, and val = P13(s,t).
truth_grid <- function(times = times_eval, s_grid = c(0,6,12),n_large=200000) {
  W <- simulate_cohort(n_large, with_censor = FALSE)

  ## True death time = min(direct death, death after progression)
  direct_time <- ifelse(W$t13_true < W$t12_true, W$t13_true, Inf)
  afterP_time <- ifelse(W$t12_true < W$t13_true, W$t23_true_abs, Inf)
  is_dead_at  <- function(t) as.numeric(direct_time <= t | afterP_time <= t)

  ## Unconditional P13(0,t)
  P3_0t <- sapply(times, function(tt) mean(is_dead_at(tt), na.rm=TRUE))
  out <- dplyr::tibble(method="TRUTH", s=0, t=times, val=P3_0t)

  ## Landmark truths: conditional on X(s)=1
```

```
  if (length(s_grid) > 0) {
    for (s in s_grid[s_grid > 0]) {
      in_s1 <- ( (is.na(W$t12_true) | W$t12_true > s) &
                   (is.na(W$t13_true) | W$t13_true > s) )
      denom <- mean(in_s1)
      vals <- if (denom > 0) {
        sapply(times[times > s], function(tt) mean(is_dead_at(tt)&in_s1)/denom)
      } else rep(NA_real_, sum(times > s))
      out <- dplyr::bind_rows(out,
          dplyr::tibble(method="TRUTH", s=s, t=times[times > s], val=vals))
    }
  }
  out
}

## 11) One replicate ----------------------------------------------------------
one_rep <- function(n = n_per_rep) {
  D <- simulate_cohort(n, with_censor = TRUE)
  dplyr::bind_rows(
    aj_from0(D, times_eval),
    landmark_aj(D, s=6,  times_eval),
    landmark_aj(D, s=12, times_eval)
  )
}

## 12) Run simulation ---------------------------------------------------------
cat("Computing Monte Carlo 'truth' (large n)...\n")
truth <- truth_grid(times = times_eval, s_grid = c(0,6,12), n_large = 200000)

cat("Running ", n_reps, " replicates with n=", n_per_rep, "...\n", sep="")
all_est <- purrr::map_dfr(seq_len(n_reps), function(i) {
  if (i %% 25 == 0) cat("  replicate ", i, "/", n_reps, "...\n", sep="")
  out <- one_rep(n_per_rep)
  out$rep <- i
  out
})

## 13) Bias and RMSE vs truth -------------------------------------------------
res <- all_est |>
  dplyr::left_join(dplyr::select(truth, s, t, val), by = c("s","t")) |>
  dplyr::filter(!is.na(val)) |>
  dplyr::mutate(err = est - val) |>
  dplyr::group_by(method, s, t) |>
  dplyr::summarise(
    n_used   = sum(!is.na(est)),
    mean_est = mean(est, na.rm = TRUE),
    truth    = dplyr::first(val),
    bias     = mean(err, na.rm = TRUE),
    rmse     = sqrt(mean(err^2, na.rm = TRUE)),
    .groups  = "drop"
  ) |>
  dplyr::arrange(s, t, method)

if (!dir.exists(outdir)) dir.create(outdir, recursive = TRUE)
readr::write_csv(all_est, file.path(outdir, "all_est.csv"))
readr::write_csv(res,     file.path(outdir, "summary_bias_rmse.csv"))

print(dplyr::glimpse(all_est))
print(res)

res <- all_est |>
```

```
  dplyr::left_join(dplyr::select(truth, s, t, val), by = c("s","t")) |>
  dplyr::filter(!is.na(val)) |>
  dplyr::group_by(method, s, t) |>
  dplyr::summarise(
    n_used   = sum(!is.na(est)),
    mean_est = mean(est, na.rm = TRUE),
    sd_est   = sd(est,   na.rm = TRUE),
    se_est   = sd_est / sqrt(n_used),
    truth    = dplyr::first(val),
    bias     = mean_est - truth,
    rmse     = sqrt(mean( (est - val)^2, na.rm = TRUE)),
    .groups  = "drop"
  ) |>
  dplyr::arrange(s, t, method)

readr::write_csv(res, file.path(outdir, "summary_bias_rmse.csv"))
print(res)

```

Appendix B. Search queries

The specific search queries employed in the Web of Science and Scopus databases are provided below. These queries were designed to capture a broad yet precise set of studies related to survival analysis in multi-state, non-Markov settings with complex censoring structures. The queries were executed on 11 June 2025 so that results may differ if the search is performed at a later date.

B.1 Web of Science query

The following query was executed using the Advanced Search mode in Web of Science, specifying the search field as "Topic" (`TS=`).

```
TS=(
("survival analysis" OR "time-to-event" OR "time to event")
AND
(
("multi-state" OR multi-state)
OR
("non-Markov" OR "non Markov")
OR
("landmark approach" OR "landmark model" OR subsampling)
)
AND
(
censor*
OR "interval censor*"
OR "left truncat*"
OR "right censor*"
OR "transition probabilit*"
OR covariat*
)
)
```

B.2 Scopus query

An equivalent query was executed in Scopus using the `TITLE-ABS-KEY` search field:

```
TITLE-ABS-KEY(
(
"survival analysis" OR "time-to-event" OR "time to event"
)
AND
(
("multi-state" OR multi-state)
OR
("non-Markov" OR "non Markov")
OR
("landmark approach" OR "landmark model" OR subsampling)
)
AND
(
censor*
OR "interval censor*"
OR "left truncat*"
OR "right censor*"
OR "transition probabilit*"
OR covariat*
)
)
```

Appendix C. Detailed overview of selected articles

Table C1.: Summary of articles on non-Markov multi-state survival analysis.

Year	Methodology	Data / Censoring
1995 [56]	Multi-state (illness–death) model for interval-censored data; application of piecewise constant incidence within a generalized linear model; development of a new graphical approach for presenting test results.	Application in a malaria study (Plasmodium falciparum) in infants; predominantly interval-censored data with some right-censoring; no simulation study reported.
1999 [50]	Multi-state semi-Markov models; discussion of key assumptions (time-homogeneity, semi-Markov, truncation, censoring); introduction of covariates via explanatory processes and penalized likelihood; general additive model for transition intensities.	Applications in epidemiology with possible truncation/censoring; observational data with incomplete records and time-dependent variables; no specific mention of a simulation study.
2000 [57]	Extension of the accelerated failure time model to multi-state processes; semiparametric approach based on estimating equations; compares treatment effects on specific sojourn times.	Application to a cancer clinical trial; potentially dependent censoring (incomplete follow-up); includes a simulation study to assess estimator performance.

Continued on the next page

Table C1 (continued from previous page)

Year	Methodology	Data / Censoring
2001 [37]	Non-parametric estimation of stage-occupation probabilities and integrated transition hazards (AJ / NA) for non-Markov multi-state systems; provides asymptotic validity proofs and tests finite-sample performance through a MC simulation study.	Independent right-censored observations; purely methodological article with extensive simulation (bias, MSE, coverage); no empirical data set analyzed.
2002 [24]	Nonparametric estimation of transition functions for non-Markov systems under dependent censoring using PCW; weighted NA form and product integration.	Application to bone marrow transplant data (graft-versus-host disease), with informative censoring; includes a simulation study for performance evaluation; dependent censoring.
2002 [38]	Flow-graph modeling of censored and truncated multi-state processes; relies on saddle-point approximations to invert Laplace transforms and includes a simulation study to assess numerical accuracy.	Simulation scenarios with right-, left-, and interval-censoring and truncation; no substantive real-world data set.

Table C2.: Summary of articles on non-Markov multi-state survival analysis

Year	Methodology	Data/Censoring
2002 [23]	Likelihood-based inference for interval-censored multi-state data using penalized piece-wise constant hazards; features a MC simulation study.	Applied to the PAQUID dementia cohort (scheduled visits → interval-censoring) and supported by simulation to examine bias and coverage.
2003 [58]	BITE software (Bayesian Intensity Estimator) for event history data analysis using flexible hierarchical models; Bayesian inference via MCMC.	Supports interval-censored or right-censored data; general approach with no specific dataset detailed; no focus on simulation (computational tool).
2004 [59]	Unified treatment of Markov and semi-Markov systems through flow-graph models; derivations use saddle-point approximations and transform inversion to obtain survival and hazard functions under censoring.	Methodological exposition with medical/engineering illustrations; no large-scale simulation, only worked numerical examples.
2004 [60]	Temporal process regression (generalized link function) for response and covariates over time; includes tests for covariate effects; does not require the Markov assumption.	Event history data with partial observation and censoring; includes a simulation study for efficiency comparison; no primary application in a specific dataset.
2006 [20]	Nonparametric estimation of transition probabilities in illness–death models without the Markov assumption; studies consistency and asymptotic normality.	Illustration with chronic disease data (right-censored); includes a simulation study and real data application; focus on non-Markov settings.
2007 [40]	`TDC.MSM`: `R` package for multi-state survival data; provides two modeling axes (regression with time-dependent covariates and five variants of multi-state models – Markov/semi-Markov, homogeneous/piecewise, and non-parametric).	Designed for multi-state data with right-censoring; includes functions to estimate transition probabilities, survival curves, and hazard functions (with graphical outputs). No dedicated simulation study is provided; focus is on implementation and illustration with real data.
2007 [39]	Landmarking approach for dynamic survival prediction fitted via pseudo-partial likelihood; contains a comprehensive simulation study investigating bias, coverage and predictive accuracy.	Right-censored follow-up with time-updated covariates; simulation scenarios and an application to renal-transplant data.

Table C3.: Overview of additional articles on non-Markov multi-state survival analysis (2007–2013).

Year	Methodology	Data/Censoring
2007 [40]	`TDC.MSM`: an R library for multi-state survival data; includes time-dependent regression models and five types of multi-state modeling (Markov or semi-Markov, homogeneous or piecewise, among others); enables Cox-based and nonparametric approaches.	Designed for right-censored multi-state datasets; provides functions for estimating transition probabilities, survival, and hazard curves; no specific simulation study highlighted; offers graphical outputs for model assessment.
2007 [61]	Asymptotic theory for the Cox semi-Markov illness–death model; derivation of large-sample properties, asymptotic variances for transition probabilities; comparison of Markov versus semi-Markov in a MC study.	Real data example from the PROVA trial (liver disease study) and additional simulation for robustness/efficiency analyses; right-censored data structure.
2009 [62]	Semi-Markov multi-state model for mean quality-adjusted survival; allows non-identically distributed sojourn times by covariate level; proposes semiparametric and parametric (exponential, Weibull) methods with jackknife variance estimation.	Contains a simulation study to evaluate performance; includes a real data set application (focus on quality-of-life outcomes); standard right-censoring framework.
2010 [41]	Flexible semi-Markov model for interval-censored data with generalized Weibull sojourn-time distributions; covariates may enter both the Markov-chain and duration components; proposes a global goodness-of-fit test.	Kidney-transplant follow-up observed only at clinic visits (interval-censoring); includes a MC simulation assessing bias, power and coverage, plus an application to transplant patients.
2012 [51]	Estimation and asymptotic theory for transition probabilities in Markov renewal (semi-Markov) multi-state models; proposes nonparametric and semiparametric estimators for forward-going models; uses functional delta-method for large-sample results.	Data from a renal replacement therapy study; right-censored observations; outlines an `R` implementation and suggests resampling methods (for example, bootstrap) for confidence bands.
2012 [63]	Nonparametric estimation of sojourn-time distributions in transient semi-Markov processes, with bootstrap-based pointwise confidence bands.	Right-censored multi-state data; includes a simulation study that investigates bootstrap coverage and finite-sample performance.
2013 [64]	Review of multi-state stochastic processes from a statistical flowgraph perspective; focuses on time-homogeneous semi-Markov models and integral transform-based algorithms; accommodates censored transition data (parametric or empirical).	Applies to various real settings (for example, reliability, medical studies) with possibly censored or incomplete data; no major emphasis on simulation; demonstrates transform-inversion methods for first passage times.
2014 [65]	Introduces a physical/electrical network analogy for multi-state survival analysis; equates the diffusion of electrical charge to changes in survival probabilities; interprets Kaplan–Meier, cause-specific Kaplan–Meier, and proportional hazards models under this circuit-based perspective.	Focus on two-state (alive–dead) models incorporating censoring and truncation; extends to competing risks with multiple sink (death) states; does not emphasize simulation, mainly conceptual/analogical framework.

Continued on next page

Table C3 (continued)

Year	Methodology	Data/Censoring
2014 [42]	Semiparametric multi-state survival methods for randomized trials; uses transform-based extrapolation beyond study follow-up and compares several extrapolation strategies.	Extensive simulation study under a range of censoring schemes, followed by an application to the LIPID cholesterol-lowering trial (right-censored).
2015 [13]	Comparative study of nonparametric estimators for transition probabilities in a non-Markov illness–death model; proposes new estimators that are consistent under general censoring assumptions (not requiring full support overlap).	Includes simulations to assess finite-sample behavior; real application to a colon cancer trial with right-censored outcomes; highlights improved efficiency versus existing non-Markov estimators.
2015 [43]	Examines the role of frailty components in multi-state models and the identifiability of unobserved heterogeneity.	Breast-cancer recurrence data (right-censoring); contains a simulation study that evaluates identifiability and statistical power when frailties are introduced.
2015 [66]	Presents a systematic procedure for specifying an illness–death model (no recovery); uses a clock-reset approach (non-homogeneous semi-Markov) and stepwise reduction to optimize predictive accuracy; tests proportional hazards assumptions on transition rates.	Illustrated with an observational study on 434 ovarian cancer patients (progression as intermediate state, death absorbing); right-censored data; no specific simulation focus.
2015 [11]	Proposes two new nonparametric estimators for transition probabilities in both progressive and non-progressive multi-state models under non-Markov assumptions; one approach generalizes difference-based Kaplan–Meier estimators, another uses constructed univariate Markov processes.	Simulation studies plus applications to colon cancer (survival and recurrence) and liver cirrhosis (prothrombin levels); right-censored data; demonstrates improved robustness and broad applicability.

Table C4.: Overview of additional articles on non-Markov multi-state survival analysis (2017–2019).

Year	Methodology	Data/Censoring
2017 [53]	Semi-Markov additive relative-survival model combining multi-state and net-survival ideas; accounts for time-in-state dependence and unknown causes of death; validated via simulation.	French kidney-transplant cohort with right-censoring; simulations show gains over standard approaches.
2017 [67]	Binomial-regression modeling of transition probabilities in a progressive illness–death model; employs randomly weighted score equations to mitigate censoring bias; coefficients interpreted as time-varying covariate effects.	Multicenter systemic-lupus registry (RELESSER); right-censoring; simulation assesses finite-sample bias and coverage.
2017 [26]	Landmark cure-rate models with time-dependent covariates; supports PH, accelerated failure time and censored-quantile specifications within a dynamic-prediction framework.	Heart-transplant data with right-censoring; MC study evaluates estimation accuracy and prediction error.
2018 [68]	multi-state recursively imputed survival trees (MSRIST) and random survival forests (MSRSF) for nonparametric estimation of transition probabilities; compares with Cox models.	Tehran HIV/AIDS cohort (2004–2014) with right-censoring; simulations (Markov and non-Markov) benchmark performance.

Continued on next page

Table C4 (continued)

Year	Methodology	Data/Censoring
2018 [55]	IPW in a semi-Markov illness–death model with interval-censoring; implementation added to the `multi-state` package of `R`.	Aortic-bioprosthesis cohort; interval-censoring; simulation confirms validity of IPW estimators under confounding.
2018 [12]	LAJ estimator for non-Markov multi-state models; retains AJ form within landmarks and shows asymptotic properties.	Series of MC experiments demonstrate small-sample bias and variance; purely methodological (no main real data).
2019 [44]	Empirical-hazard bootstrap for complex multi-state event histories without individual-patient data; simulates trajectories from aggregated hazards, handling left-truncation, right-censoring and non-Markov structures.	Proof-of-concept simulation on liver-cirrhosis progression with time-dependent treatment; demonstrates coverage and feasibility.
2019 [69]	Survey of estimation techniques for the progressive illness–death model (non-Markov); compares landmark, pre-smoothed and conditional estimators; reviews available R software.	Multiple medical examples plus simulation comparisons of estimator performance; highlights implementation details.
2019 [45]	`SemiCompRisks`: Bayesian (MCMC) and likelihood routines for independent or clustered semi-competing risks data; supports accelerated failure time/proportional hazards sub-models, flexible baseline hazards and Markov or semi-Markov specifications.	Handles right-censoring; package includes data-simulation utilities and case studies, but no large-scale MC study is emphasized.
2019 [44]	Empirical-hazard bootstrap for complex multi-state event histories without individual-patient data; simulates trajectories from an empirical multivariate hazard measure, accommodating left truncation, right censoring, non-proportional hazards and non-Markov structures; emphasizes time-dependent (qualitative) exposures and proposes empirical simulation as a novel uncertainty-assessment tool.	Approach can be driven solely by published transition-hazard information; proof-of-concept simulation study on liver-cirrhosis patients illustrates performance; handles both left truncation and right censoring within the multi-state framework.
2019 [45]	`semicompRIsks` package of `R` for semi-competing risks (illness–death model); supports independent or cluster-correlated data; flexible specification of regression models (accelerated failure time/proportional hazards), baseline hazards (parametric/nonparametric), random effects, and Markov/semi-Markov assumptions; estimation mainly Bayesian, with some maximum likelihood estimations for specific models.	Suitable for settings where a nonterminal event is subject to a terminal event; handles right-censored data (semi-competing risks framework); examples provided, but no major simulation focus.

Table C5.: Overview of additional articles on non-Markov multi-state survival analysis (2019–2025).

Year	Methodology	Data/Censoring
2019 [46]	Compares a cohort Markov model with fixed monthly transition probabilities against a partitioned-survival (non-Markov) approach that relies directly on observed survival curves; demonstrates how each framework alters incremental cost-effectiveness ratios in cervical-cancer evaluation.	Advanced-cervical-cancer case study with monthly cycles; right-censored overall- and progression-free survival curves from trial data; deterministic and probabilistic sensitivity analyses (no MC study of estimators).
2019 [52]	Derives the joint distribution of progression-free and overall survival in a three-state framework that remains valid when the Markov assumption fails; supplies closed-form estimators for the Pearson and Kendall coefficients and validates them via extensive MC simulation.	Large breast-cancer randomized controlled trial with right-censored PFS and OS; simulation assesses finite-sample bias and coverage of the correlation estimators.
2019 [70]	Landmark estimation of transition probabilities in non-Markov multi-state models with covariates; integrates subsampling at landmark times, semiparametric regression and IPW.	Population-wide Norwegian registry with work-status transitions; right-censoring; simulation shows marked differences from standard AJ under strong non-Markov violations.
2019 [71]	Estimates the Kendall coefficient between progression-free and overall survival using illness–death, copula and nonparametric approaches; compares robustness across dependence structures.	MC study plus a colon-cancer trial application; right-censoring; illness–death estimator exhibits favorable bias/variance trade-off.
2020 [47]	Introduces IPCW product-limit estimators and a local graphical test for transition probabilities when the Markov property is violated; offers practical guidance on implementation.	Simulation under various censoring/truncation scenarios and a colon-cancer data set (right-censoring); novel estimators outperform traditional AJ in non-Markov settings.
2020 [48]	Landmark dynamic-prediction model for cervical-cancer survival with time-varying covariates; updates 5-year survival probabilities and reports C-index and Brier score.	SEER cohort of 1 501 cervical-cancer cases; right-censoring; exclusively real-data evaluation (no simulation).
2021 [72]	Contrasts sojourn-time and transition-intensity parameterizations for semi-Markov multi-state models; presents estimators for both and compares them through extensive simulation as well as two real-data examples implemented in R.	Reliability and medical survival data with right-censoring; simulation examines bias and coverage under model mis-specification; vignette demonstrates practical use.
2022 [14]	Reviews inference for transition probabilities in non-Markov models; contrasts landmark, plug-in and pseudo-observation regression, and provides simulation-based guidance.	Medical illustrations (right-censoring) plus simulation comparing bias, precision and computation across estimators.
2023 [54]	Semi-Markov infectious-disease model with multiple severity levels; derives closed-form transition probabilities, waiting times and occupancy measures; extensive numerical evaluation.	Right-censored epidemic context; no MC study of estimator properties, but detailed numerical examples demonstrate applicability.

Continued on next page

Table C5 (continued)

Year	Methodology	Data/Censoring
2023 [16]	Optimal subsampling for Cox regression in massive EHR data with rare events; establishes asymptotic properties and confirms performance via simulation; applied to UK-Biobank colorectal-cancer data.	Right-censored, left-truncated EHR data with sparse events; subsampling yields major computational gains without sacrificing estimator accuracy.
2024 [27]	Tensor landmark analysis that integrates multiple longitudinal biomarkers via CP decomposition into a dynamic-prediction framework; fitted by iteratively re-weighted least squares.	MC simulation plus an ADNI Alzheimer cohort; right-censoring; tensor model improves prediction over scalar-covariate landmarking.
2025 [49]	Simulation study of a five-state semi-Markov model assessing how sample size and sojourn-time distribution affect coverage, bias and state-occupancy estimation.	Purely synthetic, right-censored data; no real application; provides practical guidance on the reliability of MSSMM inference.

Statements

Author contributions

Conceptualization: M. Azevedo; data curation: M. Azevedo; formal analysis: M. Azevedo; methodology: M. Azevedo, L. Meira-Machado, C. Moreira; supervision: L. Meira-Machado; validation: L. Meira-Machado, C. Moreira: visualization: M. Azevedo; writing original draft:M. Azevedo; writing review and editing: L. Meira-Machado, C. Moreira. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Data and code availability

The data and code used to support the findings of this study are available from the corresponding author upon request. The authors declare to honor the principles of transparency and best practice in scholarly publishing regarding data.

Declaration on the use of artificial intelligence (AI) technologies

The authors declare that no generative AI was used in the preparation of this article.

Funding

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM) Program Contract, and the project reference 2023.14897.PEX. DOI: 10.54499/2023.14897.PEX

References

[1] Therneau, T.M. and Grambsch, P.M., 2000. Modeling Survival Data: Extending the Cox Model. Springer, New York, NY, USA. DOI: 10.1007/978-1-4757-3294-8.

[2] Aalen, O.O., Borgan, Ø, and Gjessing, H.K., 2008. Survival and Event History Analysis: A Process Point of View. Statistics for Biology and Health. Springer, New York, NY, USA. DOI: 10.1007/978-0-387-68560-1.

[3] Andersen, P.K., Geskus, R.B., de Witte, T., and Putter, H., 2012. Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology, 41, 861–870. DOI: 10.1093/ije/dyr213.

[4] International Conference on Harmonization (ICH), 1998. Statistical Principles for Clinical Trials E9. ICH Harmonised Tripartite Guideline E9, CPMP/ICH/363/96. Finalized at Step 4 in February 1998. URL: https://database.ich.org/sites/default/files/E9_Guideline.pdf.

[5] ICH E9(R1) Expert Working Group, 2019. Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1). ICH Harmonised Guideline, Final version (Step 4), International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), Geneva. Adopted on 20 November 2019. URL: https://www.gmp-compliance.org/files/guidemgr/E9-R1_Step4_Guideline_2019_1203.pdf.

[6] Putter, H., Fiocco, M., and Geskus, R.B., 2007. Tutorial in biostatistics: Competing risks and multi-state models. Statistics in Medicine, 26, 2389–2430. DOI: 10.1002/sim.2712.

[7] Jackson, C., 2011. Multi-state models for panel data: The `msm` package for R. Journal of Statistical Software, 38, 1–28. DOI: 10.18637/jss.v038.i08.

[8] Zeng, D. and Lin, D.Y., 2007. Maximum likelihood estimation in semiparametric regression models with censored data. Journal of the Royal Statistical Society B, 69, 507–564. DOI: 10.1111/j.1369-7412.2007.00606.x.

[9] Geskus, R.B., 2011. Cause-specific cumulative incidence estimation and the Fine and Gray model under both left truncation and right censoring. Biometrics, 67, 39–49. DOI: 10.1111/j.1541-0420.2010.01420.x.

[10] Keiding, N., 2014. Event history analysis. Annual Review of Statistics and Its Application, 1, 333–360. DOI: 10.1146/annurev-statistics-022513-115558.

[11] Titman, A.C. and Sharples, L.D., 2015. Transition probability estimates for non-Markov multi-state models. Biometrics, 71, 1034–1041. DOI: 10.1111/biom.12349.

[12] Putter, H. and Spitoni, C., 2018. Non-parametric estimation of transition probabilities in non-Markov multi-state models: The landmark Aalen–Johansen estimator. Statistical Methods in Medical Research, 27, 2081–2092. DOI: 10.1177/0962280216674497.

[13] Meira-Machado, L. and de Uña-Álvarez, J., 2015. Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. Biometrics, 71, 364–375. DOI: 10.1111/biom.12288.

[14] Andersen, P.K., Wandall, E.N.S., and Perme, M.P., 2022. Inference for transition probabilities in non-Markov multi-state models. Lifetime Data Analysis, 28, 345–365. DOI: 10.1007/s10985-022-09560-w.

[15] Aalen, O.O. and Johansen, S., 1978. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scandinavian Journal of Statistics, 5, 141–150. URL: www.jstor.org/stable/4615704.

[16] Keret, N. and Gorfine, M., 2023. Analyzing big EHR data–optimal Cox regression subsampling procedure with rare events. Journal of the American Statistical Association, 118, 2262–2275. DOI: 10.1080/01621459.2023.2209349.

[17] Leiva, V., Castro, C., Vila, R., and Saulo, H., 2024. Unveiling patterns and trends in research on cumulative damage models for statistical and reliability analyses: Bibliometric and thematic explorations with data analytics. Chilean Journal of Statistics, 15, 81–109. DOI: 10.32372/chjs.15-01-05.

[18] Nießl, A., Allignol, A., Beyersmann, J., and Mueller, C., 2023. Statistical inference for state occupation and transition probabilities in non-Markov multi-state models subject to both random left-truncation and right-censoring. Econometrics and Statistics, 25, 110–124. DOI: 10.1016/j.ecosta.2021.09.008.

[19] Aastveit, M.E., Cunen, C., and Hjort, N.L., 2023. A new framework for semi-Markovian parametric multi-state models with interval censoring. Statistical Methods in Medical Research, 32, 1100–1123. DOI: 10.1177/09622802231160550.

[20] Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P.K., 2006. Nonparametric estimation of transition probabilities in a non-Markov illness-death model. Lifetime Data Analysis, 12, 325–344. DOI: 10.1007/s10985-006-9009-x.

[21] Horvitz, D.G. and Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663–685. DOI: 10.1080/01621459.1952.10483446.

[22] Turnbull, B.W., 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society B, 38, 290–295. DOI: j.2517-6161.1976.tb01597.x.

[23] Joly, P., Commenges, D., Helmer, C., and Letenneur, L., 2002. Inference for multistate models from interval-censored data. Statistical Methods in Medical Research, 11, 167–182. DOI: 10.1191/0962280202sm279ra.

[24] Datta, S. and Satten, G.A., 2002. Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. Biometrics, 58, 792–802. DOI: 10.1111/j.0006-341X.2002.00792.x.

[25] Andersen, P.K. and Gill, R.D., 1982. Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10, 1100–1120. URL: http://www.jstor.org/stable/2240714.

[26] Shi, H. and Yin, G., 2017. Landmark cure rate models with time-dependent covariates. Statistical Methods in Medical Research, 26, 2042–2054. DOI: 10.1177/0962280217708681.

[27] Park, S.H., Zhou, R., Zhang, X., Li, L., and Liu, L., 2024. Tensor landmark analysis with application to ADNI data. Stat, 13, e70014. DOI: 10.1002/sta4.70014.

[28] Castro, C., Leiva, V., Garrido, D., Huerta, M., and Minatogawa, V., 2024. Blockchain in clinical trials: Bibliometric and network studies of applications, challenges, and future prospects based on data analytics. Computer Methods and Programs in Biomedicine, 248, 108321. DOI: 10.1016/j.cmpb.2024.108321.

[29] Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al., 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, 71. DOI: 10.1136/bmj.n71.

[30] Pons, P. and Latapy, M., 2005. Computing communities in large networks using random walks. Yolum, P., Güngör, T., Gürgen, F., and Özturan, C. (eds.), Computer and Information Sciences, 3733, pp. 284–293. Springer, Berlin, Germany. DOI: 10.1007/11569596_31.

[31] Van Eck, N.J. and Waltman, L., 2009. How to normalize co-occurrence data? An analysis of some well-known similarity measures. Journal of the American Society for Information Science and Technology, 60, 1635–1651. DOI: 10.1002/asi.21075.

[32] Kamada, T. and Kawai, S., 1989. An algorithm for drawing general undirected graphs. Information Processing Letters, 31, 7–15. DOI: 10.1016/0020-0190(89)90102-6.

[33] Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, P10008. DOI: 10.1088/1742-5468/2008/10/P10008.

[34] Brin, S. and Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 30, 107–117. DOI: 10.1016/S0169-7552(98)00110-X.

[35] Callon, M., Courtial, J.P., and Laville, F., 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Scientometrics, 22, 155–205. DOI: 10.1007/BF02019280.

[36] Aria, M. and Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11, 959–975. DOI: 10.1016/j.joi.2017.08.007.

[37] Datta, S. and Satten, G.A., 2001. Validity of the Aalen–Johansen estimators of stage occupation probabilities and Nelson–Aalen estimators of integrated transition hazards for non-Markov models. Statistics and Probability Letters, 55, 403–411. DOI: 10.1016/S0167-7152(01)00155-9.

[38] Yau, C.L. and Huzurbazar, A.V., 2002. Analysis of censored and incomplete survival data using flowgraph models. Statistics in Medicine, 21, 3727–3743. DOI: 10.1002/sim.1237.

[39] van Houwelingen, H.C., 2007. Dynamic prediction by landmarking in event history analysis. Scandinavian Journal of Statistics, 34, 70–85. DOI: 10.1111/j.1467-9469.2006.00529.x.

[40] Meira-Machado, L., Cadarso-Suárez, C., and de Uña-Álvarez, J., 2007. TDC.MSM: An R library for the analysis of multi-state survival data. Computer Methods and Programs in Biomedicine, 86, 131–140. DOI: 10.1016/j.cmpb.2007.01.010.

[41] Foucher, Y., Giral, M., Soulillou, J.P., and Daures, J.P., 2010. A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. Statistical Methods in Medical Research, 19, 127–145. DOI: 10.1177/0962280208093889.

[42] Hudson, H.M., Lô, S.N., Simes, R.J., Tonkin, A.M., and Heritier, S., 2014. Semiparametric methods for multi-state survival models in randomised trials. Statistics in Medicine, 33, 1621–1645. DOI: 10.1002/sim.6060.

[43] Putter, H. and van Houwelingen, H.C., 2011. Frailties in multi-state models: Are they identifiable? Do we need them? Statistical Methods in Medical Research, 24, 178–197. DOI: 10.1177/0962280211424665

[44] Bluhmki, T., Putter, H., Allignol, A., and Beyersmann, J., 2019. Bootstrapping complex time-to-event data without individual patient data, with a view toward time-dependent exposures. Statistics in Medicine, 38, 3747–3763. DOI: 10.1002/sim.8177.

[45] Alvares, D., Haneuse, S., Lee, C., and Lee, K.H., 2019. SemiCompRisks: An R package for the analysis of independent and cluster-correlated semi-competing risks data. The R Journal, 11, 376–400. DOI: 10.32614/RJ-2019-038.

[46] del Campo, C., Bai, J., and Keller, L.R., 2019. Comparing Markov and non-Markov alternatives for cost-effectiveness analysis: Insights from a cervical cancer case. Operations Research for Health Care, 21, 32–43. DOI: 10.1016/j.orhc.2019.04.001.

[47] Soutinho, G. and Meira-Machado, L., 2020. Estimation of the transition probabilities in multi-state survival data: New developments and practical recommendations. WSEAS Transactions on Mathematics, 19, 353–366. DOI: 10.37394/23206.2020.19.36.

[48] Yang, Z., Hou, Y., Lyu, J., Liu, D., and Chen, Z., 2020. Dynamic prediction and prognostic analysis of patients with cervical cancer: A landmarking analysis approach. Annals of Epidemiology, 44, 45–51. DOI: 10.1016/j.annepidem.2020.01.009.

[49] Wang, Z., Li, Z., Wang, W., Li, B., Chen, W., and Ling, L., 2025. Impact of sample size, sojourn time distribution on the performance of multi-state semi-Markov model for analyzing disease progression: A simulation study. Communications in Statistics – Simulation and Computation, 54, 5427–5440. DOI: 10.1080/03610918.2025.2450708.

[50] Commenges, D., 1999. Multi-state models in epidemiology. Lifetime Data Analysis, 5, 315–327. DOI: 10.1023/A:1009636125294.

[51] Spitoni, C., Verduijn, M., and Putter, H., 2012. Estimation and asymptotic theory for transition probabilities in Markov renewal multi-state models. International Journal of Biostatistics, 8, 23. DOI: 10.1515/1557-4679.1375.

[52] Meller, M., Beyersmann, J., and Rufibach, K., 2019. Joint modeling of progression-free and overall survival and computation of correlation measures. Statistics in Medicine, 38, 4270–4289. DOI: 10.1002/sim.8295.

[53] Gillaizeau, F., Dantan, E., Giral, M., and Foucher, Y., 2015. A multi-state additive relative survival semi-Markov model. Statistical Methods in Medical Research, 26, 1700-1711. DOI: 10.1177/0962280215586456.

[54] Sukhija, S. and Kumar, R., 2023. Survival analysis of a multi-state semi-Markov model on infectious disease considering various levels of severity. Reliability: Theory and Applications, 18, 466–480. URL: cyberleninka.ru/article/n/survival-analysis-of-a-multi-state-semi-markov-model-on-infectious-disease-considering-various-levels-of-severity.

[55] Gillaizeau, F., Sénage, T., Le Borgne, F., Le Tourneau, T., Roussel, J.C., Leffondrè, K., Porcher, R., Giraudeau, B., Dantan, E., and Foucher, Y., 2017. Inverse probability weighting to control confounding in an illness–death model for interval-censored data. Statistics in Medicine, 36, 4689–4702. DOI: 10.1002/sim.7550.

[56] Andersen, P.K., Klein, J.P., and Zhang, M.J., 1995. Interval censored survival data and multi-state compartmental models in the analysis of first appearance of *Plasmodium falciparum* parasites in infants. Statistics in Medicine, 14, 2719–2732. DOI: 10.1002/sim.4780142411.

[57] Huang, Y.J., 2000. Two-sample multi-state accelerated sojourn times model. Journal of the American Statistical Association, 95, 619-627. DOI: 10.2307/2669404.

[58] Härkänen, T., 2003. BITE: A Bayesian intensity estimator. Computational Statistics, 18, 565–583. DOI: 10.1007/BF03354617.

[59] Huzurbazar, A.V., 2004. Multi-state models, flowgraph models, and semi-Markov processes. Communications in Statistics – Theory and Methods, 33, 457–474. DOI: 10.1081/STA-120028678.

[60] Fine, J.P., Yan,J., and Kosorok, M.R., 2004. Temporal process regression. Biometrika, 91, 683–703. DOI: 10.1093/biomet/91.3.683.

[61] Shu, Y., Klein, J.P., and Zhang, M.J., 2007. Asymptotic theory for the Cox semi-Markov illness-death model. Lifetime Data Analysis, 13, 91–117. DOI: 10.1007/s10985-006-9018-9.

[62] Tunes-da-Silva, G., Pedroso-de-Lima, A.C., and Sen, P.K., 2009. A semi-Markov multi-state model for estimation of the mean quality-adjusted survival for non-progressive processes. Lifetime Data Analysis, 15, 216–240. DOI: 10.1007/s10985-008-9106-0.

[63] Butler, R.W. and Bronson, D.A., 2012. Bootstrap confidence bands for sojourn distributions in multi-state semi-Markov models with right censoring. Biometrika, 99, 959–972. DOI: 10.1093/biomet/ass036.

[64] Collins, D.H. and Huzurbazar, A.V., 2013. Multi-state stochastic processes: A statistical flowgraph perspective. International Statistical Review, 81, 78–106. DOI: 10.1111/j.1751-5823.2012.00193.x.

[65] Butler, R.W. and Bronson, D.A., 2014. Multi-state survival models as transient electrical networks. Scandinavian Journal of Statistics, 41, 167–186. DOI: 10.1111/sjos.12014m

[66] Eulenburg, C., Mahner, S., Woelber, L. and Wegscheider, K., 2015. A systematic model specification procedure for an illness–death model without recovery. PLoS ONE, 10, e0123489. DOI: 10.1371/journal.pone.0123489.

[67] Azarang, L., Scheike, T., and de Uña-Álvarez, J., 2017. Direct modeling of regression effects for transition probabilities in the progressive illness–death model. Statistics in Medicine, 36, 1964–1976. DOI: 10.1002/sim.7245.

[68] Tapak, L., Kosorok, M.R., Sadeghifar, M., and Hamidi, O., 2018. Multi-state recursively imputed survival trees for time-to-event data analysis: An application to AIDS and mortality post-HIV infection data. BMC Medical Research Methodology, 18, 96. DOI: 10.1186/s12874-018-0596-5.

[69] Meira-Machado, L. and Sestelo, M., 2019. Estimation in the progressive illness–death model: A nonexhaustive review. Biometrical Journal, 61, 245–263. DOI: 10.1002/bimj.201700200.

[70] Hoff, R., Putter, H., Mehlum, I.S., Kristensen, P., and Borgan, Ø., 2019. Landmark estimation of transition probabilities in non-Markov multi-state models with covariates. Lifetime Data Analysis, 25, 660–680. DOI: 10.1007/s10985-019-09474-0.

[71] Weber, E.M., and Titman, A.C., 2019. Quantifying the association between progression-free survival and overall survival in oncology trials using Kendall's tau. Statistics in Medicine, 38, 703–719. DOI: 10.1002/sim.8001.

[72] Asanjarani, A., Liquet, B., and Nazarathy, Y., 2021. Estimation of semi-Markov multi-state models: A comparison of the sojourn times and transition intensities approaches. The International Journal of Biostatistics, 18, 243–262. DOI: 10.1515/ijb-2020-0083.

[73] Cordeiro, G.M., Biazatti, E.C., de Lima, M.C.S., and de Santana, L.H., 2023. New exponentiated–Weibull–G family: Properties, simulations, regression and applications to COVID-19 data. Chilean Journal of Statistics, 14, 99–122. DOI: 10.32372/ChJS.14-02-03.