Annals of the Institute of Statistical Mathematics (2022) 74:837–867
https://doi.org/10.1007/s10463-021-00819-x

Nonparametric tests for multistate processes with clustered data

Giorgos Bakoyannis[1] · Dipankar Bandyopadhyay[2]

Received: 25 February 2021 / Revised: 11 September 2021 / Accepted: 22 November 2021 /
Published online: 22 January 2022

Abstract
In this work, we propose nonparametric two-sample tests for population-averaged transition and state occupation probabilities for continuous-time and finite state space processes with clustered, right-censored, and/or left-truncated data. We consider settings where the two groups under comparison are independent or dependent, with or without complete cluster structure. The proposed tests do not impose assumptions regarding the structure of the within-cluster dependence and are applicable to settings with informative cluster size and/or non-Markov processes. The asymptotic properties of the tests are rigorously established using empirical process theory. Simulation studies show that the proposed tests work well even with a small number of clusters, and that they can be substantially more powerful compared to the only, to the best of our knowledge, previously proposed nonparametric test for this problem. The tests are illustrated using data from a multicenter randomized controlled trial on metastatic squamous-cell carcinoma of the head and neck.

Keywords Cluster randomized trial · Informative cluster size · Multistate model · Multicenter · Two-sample test

1 Introduction

Continuous-time stochastic processes (Capasso and Bakstein, 2015) with finite state spaces play an important role in modern medicine and public health. For example, in cancer clinical trials evaluating interventions for the underlying multistate disease processes, the patient event history often involves the states: “cancer,” “response to

✉ Giorgos Bakoyannis
gbakogia@iu.edu

[1] Department of Biostatistics and Health Data Science, Indiana University, 410 West 10th Street, Suite 3000, Indianapolis, IN 46202, USA

[2] Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, Richmond, VA 23219, USA

treatment," "disease progression," and "death." Given that "response to treatment" is an outcome that is endorsed by regulatory agencies, such as the United States Food and Drug Administration, for drug evaluation in cancer trials (US Food and Drug Administration et al., 2018), a key outcome in these trials is determining the probability of being in the "response" state as a function of time (Temkin, 1978; Begg and Larson, 1982; Ellis et al., 2008). However, response is a transient state, and its probability is not a monotonic function of time. Hence, standard methods for survival and competing risks data are not applicable for inference in these settings. When the data are independent, nonparametric estimation of transition and state occupation probabilities in general Markov processes can be performed using the Aalen–Johansen (A–J) estimator (Aalen and Johansen, 1978; Andersen et al., 2012). Datta and Satten (2001) showed that the A–J estimator is consistent for the estimation of state occupation probabilities, even for general non-Markov processes. However, estimation of transition probabilities requires appropriate extensions to this estimator (de Uña-Álvarez and Meira-Machado, 2015; Titman, 2015; Putter and Spitoni, 2018). Construction of simultaneous confidence bands and nonparametric tests for transition and state occupation probabilities can be performed using the methods by Tattar and Vaman (2014), Bluhmki et al. (2018), Bluhmki et al. (2019), and Bakoyannis (2020).

In many settings, such as in multicenter studies and cluster randomized trials (Campbell et al., 2007), the independent observations assumption is violated. Thus, the aforementioned methods are inappropriate for such settings. To the very best of our knowledge, only Bakoyannis (2021) has addressed the problem of nonparametric population-averaged estimation and two-sample testing for general multistate processes with cluster-correlated, right-censored, and/or left-truncated data. The methodology by Bakoyannis (2021) does not impose assumptions regarding the structure of the within-cluster dependence and, also, allows for informative cluster size, or ICS (Seaman et al., 2014a) scenarios, a commonplace in biomedical research, where the outcome of a cluster member is associated with the cardinality of that cluster. For the two-sample testing problem specifically, Bakoyannis (2021) proposed a nonparametric Kolmogorov–Smirnov (KS)-type test for the special case of dependent groups, where all clusters in the study include observations from both groups under comparison (complete cluster structure). However, this KS-type test may not be the most powerful for alternative hypotheses with non-crossing transition or state occupation probability functions, which is quite common in practice. Furthermore, this test is not applicable to situations where the two groups of clusters are either independent, such as in cluster randomized trials, or dependent with some clusters involving observations from one group only (incomplete cluster structure). Last but not least, a statistically significant difference based on the KS-type test does not necessarily imply that one group spends more time in a particular state (e.g., tumor response in a cancer clinical trial), which is often a hypothesis of primary interest in applications.

In this paper, we propose nonparametric two-sample tests for multistate processes with clustered, right-censored, and/or left-truncated data, and for settings with independent or dependent groups, with or without complete cluster structure. For each setting, we propose a linear test, an L^2-norm-based test, and a KS-type test. Our

testing procedures do not impose parametric assumptions and assumptions regarding the within-cluster dependence, allow for informative cluster sizes, and are applicable to both Markov and non-Markov processes. The asymptotic null distributions of the tests are established using empirical process theory (Shorack and Wellner 2009), and rigorous methodology for the calculation of p values is proposed in all cases. The L^2-norm and KS-type tests are argued to be consistent against any fixed alternative hypothesis, including alternatives with crossing transition and state occupation probability functions. Unlike the KS-type test by Bakoyannis (2021) which requires resampling methods for the calculation of p values, our linear test is asymptotically normal under the null and inference can be performed using a consistent closed-form variance estimator. Furthermore, in contrast to the KS-test, a statistically significant result based on the linear test, with a special choice of weight function, implies that one group spends more time in a particular state than the other, which is quite useful in practice. Extensive simulation studies under complex scenario show that the proposed tests work well even with a small number of clusters, and that they can be substantially more powerful compared to the test by Bakoyannis (2021) in settings with non-crossing transition and state occupation probability functions. Finally, the tests are applied to data from a multicenter randomized controlled trial on metastatic squamous-cell carcinoma of the head and neck (SCC-HN).

The structure of this paper is as follows. In Sect. 2, we introduce some notations and provide a review of the methodology for nonparametric population-averaged estimation with clustered multistate processes. In Sect. 3, we describe the proposed testing procedures along with their asymptotic properties. Sections 4 and 5 present the results from our simulation experiments and an illustration of the tests using the motivating multicenter SCC-HN data. Finally, the paper concludes with a discussion in Sect. 6. The outlines of the asymptotic theory proofs are provided in the Supplementary Materials.

2 Review of nonparametric estimation with clustered data

Let $\{X(t) : t \in [0, \tau]\}$ be a continuous-time nonhomogeneous Markov process with a finite state space $\mathcal{S} = \{1, \ldots, S\}$, absorbing state subspace $\mathcal{T} \subset \mathcal{S}$, and $\tau \in (0, \infty)$. The Markov assumption is used for simplicity of presentation here, and will be relaxed in the end of this section. If the process does not involve absorbing states, we set $\mathcal{T} = \emptyset$. The subspace of transient states $\mathcal{T}^c$ may include both non-recurrent states (e.g., for the illness-death model without recovery) and recurrent states (e.g., for the illness-death model with recovery). Let $\tilde{N}_{hj}(t)$, $h \in \mathcal{T}^c$, $j \in \mathcal{S}$, be the counting process that represents the number of direct transitions from state h to state j, with $h \neq j$, that occurred in the interval $[0, t]$. Also, let $\tilde{Y}_h(t)$, $h \in \mathcal{T}^c$, $t \in [0, \tau]$ be the indicator function with $\tilde{Y}_h(t) = 1$ if the process is at the transient state h just before time t, and $\tilde{Y}_h(t) = 0$ otherwise. The stochastic behavior of the process can be described by the $S \times S$ transition probability matrix $\mathbf{P}_0(s, t)$, $0 \leq s < t \leq \tau$, with elements $P_{0,hj}(s, t)$ defined as

$$P_{0,hj}(s,t) = P(X(t) = j | X(s) = h, \mathcal{F}_{s-}) = P(X(t) = j | X(s) = h), \quad h, j \in \mathcal{S}, \ 0 \le s < t \le \tau,$$

where $\mathcal{F}_{s-} = \sigma\langle\{\tilde{N}_{hj}(u) : 0 \le u < s, h \ne j\}\rangle$ is the history of transitions just before time s. The conditional independence of the transition probabilities from the prior history of transitions constitutes the Markov property. The stochastic behavior of the process can also be described by the transition intensities, which are defined as

$$a_{0,hj}(t) = \begin{cases} \lim_{\delta \downarrow 0} \frac{1}{\delta} P_{0,hj}(t, t+\delta) & \text{if } h \ne j \\ -\sum_{j \ne h} a_{0,hj}(t) & \text{if } h = j. \end{cases}$$

The definition of $a_{0,hj}(t)$ for the case with $h = j$ is a consequence of the fact that each row of the matrix $\mathbf{P}_0(s,t)$ is summing to 1. Another useful quantity is the cumulative transition intensity, defined as

$$A_{0,hj}(t) = \int_0^t a_{0,hj}(u)\mathrm{d}u, \quad h \in \mathcal{T}^c, \ j \in \mathcal{S}, \ t \in [0, \tau].$$

Finally, the state occupation probability for a particular state $j \in \mathcal{S}$ is defined as $P_{0,j}(t) = P(X(t) = j)$. This probability can be expressed as a function of transition probabilities:

$$P_{0,j}(t) = \sum_{h \notin \mathcal{T}^c} P_{0,h}(0) P_{0,hj}(0,t), \quad j \in \mathcal{S}, \ t \in [0, \tau].$$

In a clustered data setting, let $X_{im}(\cdot)$ be the mth process in the ith cluster, for $i = 1, \ldots, n$ and $m = 1, \ldots, M_i$. For the sake of generality, we consider the situation where cluster size M_i has the ICS property, that is M_i is associated with $X_{im}(\cdot)$. However, all methods in this manuscript are trivially applicable to simpler situations where M_i is non-informative, or fixed. Here, we consider the situation where processes from the same cluster are potentially dependent, but processes from different clusters are independent. No assumptions regarding the structure of the within-cluster dependence are imposed in this work. In practice, one observes the right-censored and (potentially) left-truncated versions of the processes $\tilde{N}_{im,hj}(t)$ and $\tilde{Y}_{im,h}(t)$, $h \in \mathcal{T}^c$, $j \in \mathcal{S}$, denoted by $N_{im,hj}(t)$ and $Y_{im,h}(t)$. $N_{im,hj}(t)$ is the number of observed direct transitions $h \to j$, with $h \ne j$, for the mth process in the ith cluster, that occurred by time t *and* before the corresponding right censoring time R_{im} and after the left truncation time L_{im}. Similarly, $Y_{im,h}(t)$ is the indicator function that the mth process in the ith cluster is at state h *and* under observation just before t. Here, we consider the situation where there is no information about transitions that occurred prior to the left truncation time, when left truncation is present. However, when left truncation is induced by cross-sectional sampling, there is such information available, and alternative methods that utilize this information are expected to be more efficient (see, e.g., de Uña-Álvarez and Mandel, 2018). The processes $\{\sum_{m=1}^{M_i} N_{im,hj}(t) : t \in [0, \tau]\}, h \ne j$, and $\{\sum_{m=1}^{M_i} Y_{im,h}(t) : t \in [0, \tau]\}, h \in \mathcal{T}^c$, are assumed to be independent and identically distributed for $i = 1, \ldots, n$. Assuming the existence of the latent processes $N_{i(M_i+1),hj}(\cdot), \ldots, N_{im_0,hj}(\cdot)$, $h \ne j$,

where m_0 is an upper bound for the cluster size (see regularity condition C2 below), and $Y_{i(M_i+1),h}(\cdot), \ldots, Y_{im_0,h}(\cdot)$, $h \in \mathcal{T}$, the latter assumption is implied if $(N_{i1,hj}(\cdot), \ldots, N_{im_0,hj}(\cdot), M_i)$, $h \neq j$, and $(Y_{i1,h}(\cdot), \ldots, Y_{im_0,h}(\cdot), M_i)$, $h \in \mathcal{T}$, are identically distributed for $i = 1, \ldots, n$, in addition to the independence assumption across clusters. The aforementioned latent processes do not contribute to our estimators or tests, but are assumed to exist for technical reasons, similarly to previous work on clustered data with random cluster sizes (see, e.g., Cai et al., 2000). These latent processes can be seen as data of potential candidate study units that could be included in the ith cluster (e.g., future patients that will attend the ith clinic). Independent and identically distributed observations assumptions across clusters are standard in the literature of statistical methods for clustered data with varying cluster sizes (see, e.g., Cai et al., 2000; Zhang et al., 2011; Liu et al., 2011; Zhou et al., 2012).

There are two populations of interest under informative cluster size: (i) the population of all cluster members (ACM) and (ii) the population of typical cluster members (TCM) (Seaman et al., 2014b; Bakoyannis, 2021). The ACM population consists of all the processes from all the clusters, while the TCM population is a subset of the ACM population consisting of a single (randomly selected) representative processes from every cluster. Clearly, larger clusters are over-represented in the ACM population, while every cluster is equally represented in the TCM population. The state occupation probability for the ACM population is defined as

$$P_{0,j}(t) = \frac{E\{M_1 I[X_{1m}(t) = j]\}}{EM_1}$$

for any $m = 1, \ldots, M_1$, where $I(\cdot)$ is the indicator function. Note that, in light of corollary 2.3.5 in Athreya and Lahiri (2006), any set function of the form $h(A) = E[MI(A)]/EM$ defined on a probability space $(\Omega, \mathcal{F}, P)$, where M is a bounded random variable with $M > 0$ almost surely, is a probability measure. Therefore, $P_{0,j}(t)$ is a well-defined probability. The state occupation probability for the TCM population is defined as $P'_{0,j}(t) = EI[X_{1m}(t) = j]$, for any $m = 1, \ldots, M_1$. The transition probabilities for the ACM population, $P_{0,hj}(s,t)$, and the TCM population, $P'_{0,hj}(s,t)$, are defined as

$$P_{0,hj}(s,t) = \frac{E\{M_1 I[X_{1m}(t) = j, X_{1m}(s) = h]\}}{E\{M_1 I[X_{1m}(s) = h]\}}, \quad h,j \in \mathcal{S}, \ \ 0 \leq s \leq t \leq \tau,$$

and

$$P'_{0,hj}(s,t) = \frac{E\{I[X_{1m}(t) = j, X_{1m}(s) = h]\}}{E\{I[X_{1m}(s) = h]\}}, \quad h,j \in \mathcal{S}, \ \ 0 \leq s \leq t \leq \tau,$$

for any $m = 1, \ldots, M_1$. The corresponding transition probability matrices $\mathbf{P}_0 = (P_{0,hj})$ and $\mathbf{P}'_0 = (P'_{0,hj})$ can also be expressed (by the Kolmogorov forward equations) as the product integrals (Bakoyannis, 2021)

$$\mathbf{P}_0(s,t) = \prod_{(s,t]} \left[\mathbf{I}_S + \mathrm{d}\mathbf{A}_0(u)\right], \quad 0 \le s \le t \le \tau,$$

where $\mathbf{A}_0$ is the cumulative transition intensity matrix for the ACM population that consists of the elements

$$A_{0,hj}(t) = \int_0^t \frac{\mathrm{d}E[M_1 \tilde{N}_{1m,hj}(u)]}{E[M_1 \tilde{Y}_{1m,h}(u)]}, \quad h \neq j,$$

and $A_{0,hh}(t) = -\sum_{h \neq j} A_{0,hj}(t)$, and

$$\mathbf{P}'_0(s,t) = \prod_{(s,t]} \left[\mathbf{I}_S + \mathrm{d}\mathbf{A}'_0(u)\right], \quad 0 \le s \le t \le \tau,$$

where $\mathbf{A}'_0$ is the cumulative transition intensity matrix for the TCM population that consists of the elements

$$A'_{0,hj}(t) = \int_0^t \frac{\mathrm{d}E[\tilde{N}_{1m,hj}(u)]}{E[\tilde{Y}_{1m,h}(u)]}, \quad h \neq j,$$

and $A'_{0,hh}(t) = -\sum_{h \neq j} A'_{0,hj}(t)$.

The transition probability matrix $\mathbf{P}_0$ for the ACM population can be estimated nonparametrically using the moment-based estimator (Bakoyannis, 2021)

$$\hat{\mathbf{P}}_n(s,t) = \prod_{(s,t]} \left[\mathbf{I}_S + \mathrm{d}\hat{\mathbf{A}}_n(u)\right], \quad 0 \le s \le t \le \tau,$$

where $\prod$ denotes the product integral, $\mathbf{I}_S$ is the $S \times S$ identity matrix, and $\hat{\mathbf{A}}_n(t)$ is the matrix consisting of the elements

$$\hat{A}_{n,hj}(t) = \int_0^t \frac{\sum_{i=1}^n \sum_{m=1}^{M_i} \mathrm{d}N_{im,hj}(u)}{\sum_{i=1}^n \sum_{m=1}^{M_i} Y_{im,h}(u)}, \quad h \neq j,$$

and $\hat{A}_{n,hh}(t) = -\sum_{j \neq h} \hat{A}_{n,hj}(t)$. The transition probability matrix $\mathbf{P}'_0$ for the TCM population can be estimated nonparametrically using the weighted moment-based estimator (Bakoyannis, 2021)

$$\hat{\mathbf{P}}'_n(s,t) = \prod_{(s,t]} \left[\mathbf{I}_S + \mathrm{d}\hat{\mathbf{A}}'_n(u)\right], \quad 0 \le s \le t \le \tau,$$

where $\hat{\mathbf{A}}'_n(t)$ is the matrix consisting of the elements

$$\hat{A}'_{n,hj}(t) = \int_0^t \frac{\sum_{i=1}^n M_i^{-1} \sum_{m=1}^{M_i} \mathrm{d}N_{im,hj}(u)}{\sum_{i=1}^n M_i^{-1} \sum_{m=1}^{M_i} Y_{im,h}(u)}, \quad h \neq j,$$

and $\hat{A}'_{n,hh}(t) = -\sum_{j \neq h} \hat{A}'_{n,hj}(t)$. Based on the estimated transition probabilities, the nonparametric estimators of the state occupation probabilities for the ACM and TCM populations are (Bakoyannis, 2021)

$$\hat{P}_{n,j}(t) = \sum_{h \in \mathcal{T}^c} \left[\frac{\sum_{i=1}^n \sum_{m=1}^{M_i} Y_{im,h}(0+)}{\hat{\pi}_n \sum_{i=1}^n M_i} \right] \hat{P}_{n,hj}(0,t),$$

where $\hat{\pi}_n = n^{-1} \sum_{i=1}^n M_i^{-1} \sum_{m=1}^{M_i} \sum_{h \in \mathcal{T}^c} Y_{im,h}(0+)$, and

$$\hat{P}'_{n,j}(t) = \sum_{h \in \mathcal{T}^c} \left[\frac{\sum_{i=1}^n \frac{1}{M_i} \sum_{m=1}^{M_i} Y_{im,h}(0+)}{n\hat{\pi}_n} \right] \hat{P}'_{n,hj}(0,t),$$

respectively. Note that $\hat{\pi}_n$ is an estimator of the probability π_0 that a process is not left-truncated. Clearly, the state occupation probability estimators are valid, only if π_0 is bounded away from zero, i.e., in settings where not all observations are left-truncated. If there is no left truncation, then $\hat{\pi}_n = 1$. It is important to note that $\hat{P}_{n,j}(t)$ and $\hat{P}'_{n,j}(t)$ are uniformly consistent, even if the process $\{X(t) : t \in [0, \tau]\}$ is not Markov (Datta and Satten, 2001; Bakoyannis, 2021). However, this is not true for $\hat{\mathbf{P}}_n(s,t)$ and $\hat{\mathbf{P}}'_n(s,t)$, when $s > 0$. In non-Markov settings, one can instead use the landmark versions of $\hat{\mathbf{P}}_n(s,t)$ and $\hat{\mathbf{P}}'_n(s,t)$ (Putter and Spitoni, 2018; Bakoyannis, 2021), which can be obtained by imposing a simple modification in $N_{im,hj}(t)$ and $Y_{im,h}(t)$ (for more details see Bakoyannis, 2021).

From a practical standpoint, selecting the most appropriate target population in situations with random and informative cluster size requires a careful consideration of the scientific goal of the study. As an example, consider a multicenter clinical trial on metastatic squamous-cell carcinoma of the head and neck with the goal of comparing the efficacy of the combination of chemotherapy and panitumumab versus chemotherapy alone. In such a trial, evaluating the effect of the combined treatment on the population-averaged probability of tumor response over the ACM population (i.e., the population of all clinic patients) is more relevant if the goal is to understand the effect of treatment on a *typical patient* from the population of all patients in all clinics. This analysis could provide evidence to regulatory agencies for broad public policy decisions and recommendations regarding the approval of the combined treatment. On the other hand, the population-averaged probability of response over the TCM population (i.e., the population of typical clinic patients) is more relevant for understanding the average treatment effect on a typical patient from a *typical clinic setting*. This could be beneficial from a health services research perspective in order to study the effectiveness of the combined treatment in the average-performing clinic. In addition, evaluating the population-averaged probability of death over the TCM population, would provide evidence for the burden of death from the particular disease in the average clinic. From this example, it is evident that

if the unit of main scientific interest is the cluster member, then inference about the ACM population is more relevant. In contrast, if the cluster is the main unit of interest, then inference about the TCM population is more desirable.

3 Nonparametric two-sample testing

In this section, we address the problem of comparing transition and state occupation probabilities for a particular transition $h \to j$ of the process $X(t)$ between two groups, say groups 1 and 2. Depending on what is the most scientifically relevant population-averaged quantity in a given setting, the null hypothesis is either $H_0 : P_{0,1hj}(s,\cdot) = P_{0,2hj}(s,\cdot)$, or $H_0 : P'_{0,1hj}(s,\cdot) = P'_{0,2hj}(s,\cdot)$, for some $s \in [0,\tau)$. The corresponding two-sided alternative hypotheses are $H_1 : P_{0,1hj}(s,\cdot) \neq P_{0,2hj}(s,\cdot)$ and $H_1 : P'_{0,1hj}(s,\cdot) \neq P'_{0,2hj}(s,\cdot)$. Alternatively, one may be interested in comparing the state occupation probabilities for a particular state $j \in \mathcal{S}$ between the two groups. The null hypothesis in this case is either $H_0 : P_{0,1j} = P_{0,2j}$, or $H_0 : P'_{0,1j} = P'_{0,2j}$. Testing such hypotheses can be based on a sample of clusters of observations of the process of interest, which satisfies the requirements described in Sect. 2. Here, we denote the counting and indicator processes for the mth observation in the pth group in the ith cluster as $N_{ipm,hj}(t)$, $h \neq j$, and $Y_{ipm,h}(t)$, $h \in \mathcal{T}^c$. We also denote the probability that a process in group p is not left-truncated by $\pi_{0,p}$, the number of observations in the ith cluster in the pth group by M_{ip}, and the transition intensities for the ACM and TCM populations for group p as $A_{0,phj}(t)$ and $A'_{0,phj}(t)$, respectively. Finally, we define the following influence functions that appear in the asymptotic null distributions of the proposed test statistics:

$$\gamma_{i,phj}(s,t) = \sum_{l\in\mathcal{T}^c}\sum_{q\in\mathcal{S}}\sum_{m=1}^{M_{ip}}\int_s^t \frac{P_{0,phl}(s,u-)P_{0,pqj}(u,t)}{E\left[\sum_{m=1}^{M_{1p}} Y_{1pm,l}(u)\right]}\mathrm{d}U_{ipm,lq}(u), \quad h \neq j,$$

where $U_{ipm,lq}(t) = N_{ipm,lq}(t) - \int_{(0,t]} Y_{ipm,l}(u)\mathrm{d}A_{0,plq}(u)$ and $\gamma_{i,phh}(s,t) = -\sum_{j\neq h}\gamma_{i,phj}(s,t)$,

$$\gamma'_{i,phj}(s,t) = \sum_{l\in\mathcal{T}^c}\sum_{q\in\mathcal{S}}\frac{1}{M_{ip}}\sum_{m=1}^{M_{ip}}\int_s^t \frac{P'_{0,phl}(s,u-)P'_{0,pqj}(u,t)}{E\left[M_{1p}^{-1}\sum_{m=1}^{M_1} Y_{1pm,l}(u)\right]}\mathrm{d}U'_{ipm,lq}(u), \quad h \neq j,$$

where $U'_{ipm,lq}(t) = N_{ipm,lq}(t) - \int_{(0,t]} Y_{ipm,l}(u)\mathrm{d}A'_{0,plq}(u)$ and $\gamma'_{i,phh}(s,t) = -\sum_{j\neq h}\gamma'_{i,phj}(s,t)$,

$$\psi_{i,pj}(t) = \sum_{h\in\mathcal{T}^\tau} \Bigg(P_{0,ph}(0)\gamma_{i,phj}(0,t) + P_{0,phj}(0,t)\Bigg\{ \sum_{m=1}^{M_{ip}} \frac{Y_{ipm,h}(0+) - EY_{1pm,h}(0+)}{\pi_{0,p} EM_{1p}} - P_{0,ph}(0)\left[\frac{M_i - EM_{1p}}{EM_{1p}} + \frac{\sum_{h\in\mathcal{T}^\tau} M_{ip}^{-1} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(0+) - \pi_{0,p}}{\pi_{0,p}}\right]\Bigg\}\Bigg),$$

and

$$\psi'_{i,pj}(t) = \sum_{h\in\mathcal{T}^\tau} \Bigg(P'_{0,ph}(0)\gamma'_{i,phj}(0,t) + \frac{P'_{0,phj}(0,t)}{\pi_{0,p}}\Bigg\{ \frac{1}{M_{ip}} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(0+) - E\left[\frac{1}{M_{1p}} \sum_{m=1}^{M_{1p}} Y_{1pm,h}(0+)\right] - P'_{0,ph}(0)\left[\sum_{h\in\mathcal{T}^\tau} \frac{1}{M_{ip}} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(0+) - \pi_{0,p}\right]\Bigg\}\Bigg).$$

The next subsections present appropriate hypothesis testing procedures for (i) independent groups and (ii) dependent groups. For simplicity of presentation, we consider the case where the process $\{X(t) : t \in [0, \tau]\}$ is Markov. However, the inference procedures presented here are also applicable to non-Markov processes, with the exception of using the landmark versions of the transition probability estimators and the landmark versions of $N_{ipm,hj}(t)$ and $Y_{ipm,h}(t)$, when $s > 0$, in the testing procedures for $P_{0,phj}(s,t)$ and $P'_{0,phj}(s,t)$ (for more details see Bakoyannis, 2021).

3.1 Independent groups

In the independent-groups case, one observes two groups of clusters, with sizes n_1 and n_2. An example of this situation is a cluster randomized trial where a new intervention is applied to a random group of clusters (e.g., clinics) only, while standard of care is used in the remaining clusters of the study. Based on two independent groups of clusters, the estimators of the pointwise between-group difference with respect to the population-averaged transition probabilities are defined as

$$\hat{\Delta}_{n_1,n_2,hj}(s,t) = \left[\hat{P}_{n_1,1hj}(s,t) - \hat{P}_{n_2,2hj}(s,t)\right], \quad t \in [s,\tau],$$

where $\hat{P}_{n_p,phj}$, $p = 1, 2$, is the estimator of $P_{0,phj}$ from the pth group, and

$$\hat{\Delta}'_{n_1,n_2,hj}(s,t) = \left[\hat{P}'_{n_1,1hj}(s,t) - \hat{P}'_{n_2,2hj}(s,t)\right], \quad t \in [s,\tau],$$

where $\hat{P}'_{n_p,phj}$, $p = 1, 2$, is the estimator of $P'_{0,phj}$ from the pth group, for some $s \in [0, \tau)$. Similarly, define the differences between the population-averaged state occupation probabilities as

$$\hat{\Delta}_{n_1,n_2,j}(t) = \left[\hat{P}_{n_1,1j}(t) - \hat{P}_{n_2,2j}(t)\right], \quad t \in [0, \tau],$$

where $\hat{P}_{n_p,pj}$, $p = 1, 2$, is the estimator of $P_{0,pj}$ from the pth group, and

$$\hat{\Delta}'_{n_1,n_2,j}(t) = \left[\hat{P}'_{n_1,1j}(t) - \hat{P}'_{n_2,2j}(t)\right], \quad t \in [0, \tau],$$

where $\hat{P}'_{n_p,pj}$, $p = 1, 2$, is the estimator of $P'_{0,pj}$ from the pth group. The corresponding nonparametric cluster bootstrap realizations of the above differences are denoted by $\hat{\Delta}^*_{n_1,n_2,hj}(s,t)$, $\hat{\Delta}'^*_{n_1,n_2,hj}(s,t)$, $\hat{\Delta}^*_{n_1,n_2,j}(t)$, and $\hat{\Delta}'^*_{n_1,n_2,j}(t)$. These can be calculated by randomly sampling clusters with replacement, and calculating the desired estimator using the resulting bootstrap dataset. Explicit expressions for the bootstrap versions of the estimators are provided in the proof of Theorem 2 in the Supplementary Material. Based on these differences, we define the following linear test statistics:

$$Z_{n_1,n_2,hj}(s) = \int_s^\tau \hat{W}_{hj}(t)\hat{\Delta}_{n_1,n_2,hj}(s,t)\mathrm{d}\mu(t), \quad \text{for some } s \in [0, \tau),$$

$$Z'_{n_1,n_2,hj}(s) = \int_s^\tau \hat{W}'_{hj}(t)\hat{\Delta}'_{n_1,n_2,hj}(s,t)\mathrm{d}\mu(t), \quad \text{for some } s \in [0, \tau),$$

$$Z_{n_1,n_2,j} = \int_0^\tau \hat{W}_j(t)\hat{\Delta}_{n_1,n_2,j}(t)\mathrm{d}\mu(t),$$

and

$$Z'_{n_1,n_2,j} = \int_0^\tau \hat{W}'_j(t)\hat{\Delta}'_{n_1,n_2,j}(t)\mathrm{d}\mu(t),$$

where $\hat{W}_{hj}(t)$, $\hat{W}'_{hj}(t)$, $\hat{W}_j(t)$ and $\hat{W}'_j(t)$ are appropriate weight functions (see condition C7 below), and the integrator $\mu(t) = t$ induces the Lebesgue measure defined on the Borel σ-algebra on $[0, \tau]$. Essentially, these linear test statistics represent the areas under the weighted difference curves $\hat{\Delta}_{n_1,n_2,hj}(s,\cdot)$, $\hat{\Delta}'_{n_1,n_2,hj}(s,\cdot)$, $\hat{\Delta}_{n_1,n_2,j}$, or $\hat{\Delta}'_{n_1,n_2,j}$. In particular, the test statistics $Z_{n_1,n_2,j}$ and $Z'_{n_1,n_2,j}$ with the weight functions $\hat{W}_j(t) = \hat{W}'_j(t) = 1$ represent the between-group difference in state-specific life expectancy. The importance of the weight functions lies on the fact that they can restrict the comparison interval to a set of times where both groups under comparison have nonzero observations at risk for the transition of interest. An example of such a weight function is

$$\hat{W}_{hj}(t) = I\left[\prod_{l \in L(h,j)} \bar{Y}_{1,l}(t)\bar{Y}_{2,l}(t) > 0\right],$$

where $L(h,j) = \{d \in \mathcal{S} : d$ is a transient state that can be visited during the transition $h \to j\}$ and $\bar{Y}_{p,h}(t) = n_p^{-1} \sum_{i=1}^{n_p} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(t)$, for the group $p = 1, 2$. Similarly, this type of weight can be defined for the state occupation probabilities as

$$\hat{W}_j(t) = I\left[\prod_{l \in \cup_{h \in \mathcal{T}^c} L(h,j)} \bar{Y}_{1,l}(t)\bar{Y}_{2,l}(t) > 0\right].$$

The weights $\hat{W}'_{hj}(t)$ and $\hat{W}'_j(t)$ are defined similarly by replacing $\bar{Y}_{p,h}(t)$ with $n_p^{-1} \sum_{i=1}^{n_p} M_{ip}^{-1} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(t)$, $p = 1, 2$. The weight functions can also be used to assign less weight to observation times with a smaller number of observations at risk, where the estimated difference tends to be unstable. An example of such weight function is

$$\hat{W}_{hj}(t) = \frac{\prod_{l \in L(h,j)} \bar{Y}_{1,l}(t)\bar{Y}_{2,l}(t)}{\sum_{l \in L(h,j)} \left[\bar{Y}_{1,l}(t) + \bar{Y}_{2,l}(t)\right]},$$

and

$$\hat{W}_j(t) = \frac{\prod_{l \in \cup_{h \in \mathcal{T}^c} L(h,j)} \bar{Y}_{1,l}(t)\bar{Y}_{2,l}(t)}{\sum_{l \in \cup_{h \in \mathcal{T}^c} L(h,j)} \left[\bar{Y}_{1,l}(t) + \bar{Y}_{2,l}(t)\right]}.$$

The corresponding weights $\hat{W}'_{hj}(t)$ and $\hat{W}'_j(t)$ can be defined similarly by replacing $\bar{Y}_{p,h}(t)$ with $n_p^{-1} \sum_{i=1}^{n_p} M_{ip}^{-1} \sum_{m=1}^{M_{ip}} Y_{ipm,h}(t)$, $p = 1, 2$. In practice, we suggest the use of this latter type of weight functions.

In what follows, we assume the following regularity conditions:

C1. The potential left truncation L_{ipm} and right censoring R_{ipm} times are independent of the underlying counting processes $\{\tilde{N}_{ipm,hj}(t) : h \neq j, t \in [0, \tau]\}$, the initial state indicators $\tilde{Y}_{ipm,h}(0+)$, $h \in \mathcal{T}^c$, and the cluster size M_{ip}. Also, L_{ipm} and R_{ipm} are identically distributed in the sense that $E[\{I(L_{ipm,1} = 0) + I(L_{ipm} < t)\}I(R_{ipm} \geq t)] = E[\{I(L_{ip1} = 0) + I(L_{ip1} < t)\}I(R_{ip1} \geq t)]$, $t \in [0, \tau]$, for any $i = 1, \ldots, n_p$, $p = 1, 2$, and $m = 1, \ldots, M_{ip}$.

C2. The cluster size is bounded in the sense that there exists a (fixed) positive integer m_0 such that $P(M > m_0) = 0$.

C3. The underlying counting processes are identically distributed conditionally on cluster size, which implies that $E\{\tilde{N}_{ipm,hj}(t)|M_{ip}\} = E\{\tilde{N}_{ip1,hj}(t)|M_{ip}\}$ for any $m = 1, \ldots, M_{ip}$ and $h \neq j$. Also, $E\{\tilde{N}_{ipm,hj}(\tau)\}^2 < \infty$ for all $h \neq j$.

C4. The underlying at-risk processes are identically distributed conditionally on cluster size, which implies that $E\{\tilde{Y}_{ipm,h}(t)|M_{ip}\} = E\{\tilde{Y}_{ip1,h}(t)|M_{ip}\}$ for any $i = 1, \ldots, n_p$, $p = 1, 2$, $m = 1, \ldots, M_{ip}$ and $h \in \mathcal{S}$. Also, there exists a convex and compact set $J_h \subset [0, \tau]$ such that $\inf_{t \in J_h} E\{\sum_{m=1}^{M_{ip}} Y_{ipm,h}(t)\} > 0$ for all $h \in \mathcal{T}^c$, and $\int_{(0,t] \cap J_h^c} dA_{0,hj}(t) = 0$ for all $h \in \mathcal{T}^c$ and $j \neq h$.

C5. The cumulative transition intensities $\{A_{0,phj}(t) : p = 1, 2, h \neq j, t \in [0, \tau]\}$ and $\{A'_{0,phj}(t) : p = 1, 2, h \neq j, t \in [0, \tau]\}$ are continuous functions.
C6. Strengthen condition C4 to require $\inf_{t\in[0,\tau]} E\{\sum_{m=1}^{M_{ip}} Y_{ipm,h}(t)\} > 0$ for all $h \in \mathcal{T}^c$.
C7. The weight functions $\hat{W}_{hj}(t)$, $\hat{W}'_{hj}(t)$, $\hat{W}_j(t)$ and $\hat{W}'_j(t)$ are uniformly consistent (in probability) for the non-negative, uniformly bounded, and cadlag fixed functions $W_{hj}(t)$, $W'_{hj}(t)$, $W_j(t)$ and $W'_j(t)$.

It is important to note that the weight functions introduced earlier satisfy condition C7. This follows from the fact that the classes of functions $\{\sum_{m=1}^{M_p} Y_{pm,h}(t) : t \in [0, \tau]\}$ and $\{M_p^{-1} \sum_{m=1}^{M_p} Y_{pm,h}(t) : t \in [0, \tau]\}$, $p = 1, 2$, $h \in \mathcal{T}^c$, are P-Donsker in light of conditions C2 and C3 (see Web Appendix in Bakoyannis, 2021), which implies that these classes are also P-Glivenko–Cantelli, conditions C5 and C6, and the continuous mapping theorem (Kosorok, 2008).

Theorem 1 states the asymptotic null distributions of the linear tests as $n_1 \wedge n_2 \to \infty$. In what follows, weak convergence is denoted by $\rightsquigarrow$.

Theorem 1 *Suppose that conditions C1–C7 hold and that* $n_1/(n_1 + n_2) \to \lambda \in (0, 1)$ *as* $n_1 \wedge n_2 \to \infty$. *Then, under the null hypothesis and for any* $h \in \mathcal{T}^c$, $j \in \mathcal{S}$, *and* $s \in [0, \tau)$, *the following hold*

(i) $\sqrt{\frac{n_1 n_2}{n_1+n_2}} Z_{n_1,n_2,hj}(s) \rightsquigarrow G_{hj}(s)$ *as* $n_1 \wedge n_2 \to \infty$, *where* $G_{hj}(s) \sim N(0, \omega^2_{hj}(s))$ *with*

$$\omega^2_{hj}(s) = (1-\lambda)E\left[\int_s^\tau W_{hj}(t)\gamma_{1,1hj}(s,t)\mathrm{d}\mu(t)\right]^2 + \lambda E\left[\int_s^\tau W_{hj}(t)\gamma_{1,2hj}(s,t)\mathrm{d}\mu(t)\right]^2.$$

(ii) $\sqrt{\frac{n_1 n_2}{n_1+n_2}} Z_{n_1,n_2,j} \rightsquigarrow G_j$ *as* $n_1 \wedge n_2 \to \infty$, *where* $G_j \sim N(0, \omega^2_j)$ *with*

$$\omega^2_j = (1-\lambda)E\left[\int_0^\tau W_j(t)\psi_{1,1j}(t)\mathrm{d}\mu(t)\right]^2 + \lambda E\left[\int_0^\tau W_{hj}(t)\psi_{1,2j}(t)\mathrm{d}\mu(t)\right]^2.$$

The proof of Theorem 1 is provided in the Supplementary Material. Consistent (in probability) estimators of the asymptotic variances in Theorem 1 are

$$\hat{\omega}^2_{hj}(s) = \frac{n_2}{(n_1+n_2)n_1}\sum_{i=1}^{n_1}\left[\int_s^\tau \hat{W}_{hj}(t)\hat{\gamma}_{i,1hj}(s,t)\mathrm{d}\mu(t)\right]^2 + \frac{n_1}{(n_1+n_2)n_2}\sum_{i=1}^{n_2}\left[\int_s^\tau \hat{W}_{hj}(t)\hat{\gamma}_{i,2hj}(s,t)\mathrm{d}\mu(t)\right]^2,$$

and

$$\hat{\omega}_j^2 = \frac{n_2}{(n_1+n_2)n_1}\sum_{i=1}^{n_1}\left[\int_0^\tau \hat{W}_j(t)\hat{\psi}_{i,1j}(t)\mathrm{d}\mu(t)\right]^2 + \frac{n_1}{(n_1+n_2)n_2}\sum_{i=1}^{n_2}\left[\int_0^\tau \hat{W}_{hj}(t)\hat{\psi}_{i,2j}(t)\mathrm{d}\mu(t)\right]^2,$$

where $\hat{\gamma}_{i,phj}(s,t)$ and $\hat{\psi}_{i,pj}(t)$ are the empirical versions of the influence functions $\gamma_{i,phj}(s,t)$ and $\psi_{i,pj}(t)$, $p=1,2$. These empirical versions can be obtained by replacing unknown parameters with their consistent estimates and expectations with sample averages over clusters. Alternatively, by Theorem 2 in Bakoyannis (2021) and the bootstrap continuous mapping theorem (Theorem 10.8 in Kosorok, 2008), these variances can be estimated as sample variances based on a number of nonparametric cluster bootstrap realizations $\int_s^\tau \hat{W}_{hj}(t)[\hat{\Delta}^*_{n_1,n_2,hj}(s,t)-\hat{\Delta}_{n_1,n_2,hj}(s,t)]\mathrm{d}\mu(t)$ and $\int_0^\tau \hat{W}_j(t)[\hat{\Delta}^*_{n_1,n_2,j}(t)-\hat{\Delta}_{n_1,n_2,j}(t)]\mathrm{d}\mu(t)$, respectively. Based on any of these variance estimators and Theorem 1, it is easy to construct an asymptotic Z-test for testing the null hypothesis of interest as usual. Using the same arguments as those used in the proof of Theorem 1, it can also be shown that a similar version of this theorem holds for the test statistics $Z'_{n_1,n_2,hj}(s)$ and $Z'_{n_1,n_2,j}$.

Even though the linear tests are expected to have a good power for alternatives with non-crossing probability functions, these tests may not be the best choice for situations where the two probability functions under comparison cross at one or more time points. For such situations, we propose the L^2-norm-based tests

$$Q_{n_1,n_2,hj}(s) = \left\{\int_s^\tau \left[\hat{W}_{hj}(t)\hat{\Delta}_{n_1,n_2,hj}(s,t)\right]^2\mathrm{d}\mu(t)\right\}^{1/2}, \quad \text{for some } s\in[0,\tau),$$

and

$$Q_{n_1,n_2,j} = \left\{\int_0^\tau \left[\hat{W}_j(t)\hat{\Delta}_{n_1,n_2,j}(t)\right]^2\mathrm{d}\mu(t)\right\}^{1/2},$$

and the KS-type tests

$$K_{n_1,n_2,hj}(s) = \sup_{t\in[s,\tau]}\left|\hat{W}_{hj}(t)\hat{\Delta}_{n_1,n_2,hj}(s,t)\right|, \quad \text{for some } s\in[0,\tau),$$

and

$$K_{n_1,n_2,j} = \sup_{t\in[0,\tau]}\left|\hat{W}_j(t)\hat{\Delta}_{n_1,n_2,j}(t)\right|.$$

The corresponding tests for $\hat{\Delta}'_{n_1,n_2,hj}(s,t)$ and $\hat{\Delta}'_{n_1,n_2,j}(t)$, denoted by $Q'_{n_1,n_2,hj}(s)$, $K'_{n_1,n_2,hj}(s)$, and $Q'_{n_1,n_2,j}$, $K'_{n_1,n_2,j}$, are defined in the same manner. We must note that the KS-type tests have the same structure as that in Bakoyannis (2021), however, with different asymptotic null distributions. Unlike the linear tests, the L^2-norm and the KS-type tests are not asymptotically normal under the null hypothesis.

Conducting hypothesis testing with these tests can be based on a resampling technique that utilizes the estimated *multiplier* processes

$$\hat{C}_{n_1,n_2,hj}(s,t) = \sqrt{\frac{n_2}{n_1+n_2}}\hat{W}_{hj}(t)\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{\gamma}_{i,1hj}(s,t)\xi_{i1} - \sqrt{\frac{n_1}{n_1+n_2}}\hat{W}_{hj}(t)\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\hat{\gamma}_{i,2hj}(s,t)\xi_{i2}, \quad t \in [s,\tau],$$

for some $s \in [0,\tau)$, where ξ_{ip}, $p = 1,2$, $i = 1,\ldots,n_p$, are independent standard normal variables, and

$$\hat{C}_{n_1,n_2,j}(t) = \sqrt{\frac{n_2}{n_1+n_2}}\hat{W}_{j}(t)\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{\psi}_{i,1j}(t)\xi_{i1} - \sqrt{\frac{n_1}{n_1+n_2}}\hat{W}_{j}(t)\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\hat{\psi}_{i,2j}(t)\xi_{i2}, \quad t \in [0,\tau].$$

Similarly, one can define the estimated multiplier processes $\hat{C}_{n_1,n_2,hj}(s,t)$ and $\hat{C}'_{n_1,n_2,j}(t)$ which correspond to the tests for $\hat{\Delta}'_{n_1,n_2,hj}(s,t)$ and $\hat{\Delta}'_{n_1,n_2,j}(t)$. Alternatively, one can use the nonparametric cluster bootstrap (Cameron et al., 2008) for inference. Theorem 2 provides the basis for conducting hypothesis testing based on the L^2-norm-based and KS-type tests.

Theorem 2 *Suppose that conditions C1–C7 hold and that* $n_1/(n_1+n_2) \to \lambda \in (0,1)$ *as* $n_1 \wedge n_2 \to \infty$. *Then, under the null hypothesis and for any* $h \in \mathcal{T}^c$, $j \in \mathcal{S}$, *and some* $s \in [0,\tau)$, *the following hold*

(i) $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\hat{W}_{hj}(\cdot)\hat{\Delta}_{n_1,n_2,hj}(s,\cdot) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1hj}(s,\cdot) - \sqrt{\lambda}\mathbb{G}_{2hj}(s,\cdot)$ *in* $D[s,\tau]$ *as* $n_1 \wedge n_2 \to \infty$, *where* $\mathbb{G}_{phj}(s,\cdot)$, $p = 1,2$, *are two independent tight zero-mean Gaussian processes with covariance functions* $W_{hj}(t_1)W_{hj}(t_2)E[\gamma_{1,phj}(s,t_1)\gamma_{1,phj}(s,t_2)]$, *for* $t_1, t_2 \in [s,\tau]$. *Moreover,*

$$\hat{C}_{n_1,n_2,hj}(s,\cdot) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1hj}(s,\cdot) - \sqrt{\lambda}\mathbb{G}_{2hj}(s,\cdot) \text{ in } D[s,\tau],$$

conditionally on the observed data and

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\hat{W}_{hj}(\cdot)[\hat{\Delta}^*_{n_1,n_2,hj}(s,\cdot) - \hat{\Delta}_{n_1,n_2,hj}(s,\cdot)] \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1hj}(s,\cdot) - \sqrt{\lambda}\mathbb{G}_{2hj}(s,\cdot) \text{ in } D[s,\tau],$$

conditionally on the observed data.

(ii) $\sqrt{\frac{n_1 n_2}{n_1+n_2}}\hat{W}_j\hat{\Delta}_{n_1,n_2,j} \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1j} - \sqrt{\lambda}\mathbb{G}_{2j}$ *in* $D[0,\tau]$ *as* $n_1 \wedge n_2 \to \infty$, *where* $\mathbb{G}_{pj}$, $p = 1, 2$, *are two independent tight zero-mean Gaussian processes with covariance functions* $W_j(t_1)W_j(t_2)E[\psi_{1,pj}(t_1)\psi_{1,pj}(t_2)]$, *for* $t_1, t_2 \in [0,\tau]$. *Moreover*,

$$\hat{C}_{n_1,n_2,j} \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1j} - \sqrt{\lambda}\mathbb{G}_{2j} \text{ in } D[0,\tau],$$

conditionally on the observed data and

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\hat{W}_j(\hat{\Delta}^*_{n_1,n_2,j} - \hat{\Delta}_{n_1,n_2,j}) \rightsquigarrow \sqrt{1-\lambda}\mathbb{G}_{1j} - \sqrt{\lambda}\mathbb{G}_{2j} \text{ in } D[0,\tau],$$

conditionally on the observed data.

The proof of Theorem 2 is provided in the Supplementary Material. Using the same arguments to those used in the proof of Theorem 2, it can be easily shown that a similar version of this theorem holds for the differences $\hat{\Delta}'_{n_1,n_2,hj}(s,\cdot)$ and $\hat{\Delta}'_{n_1,n_2,j}$. Theorem 2 along with the continuous mapping theorem leads to following asymptotic null distributions of the test statistics:

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}Q_{n_1,n_2,hj}(s) \rightsquigarrow \left\{\int_s^\tau \left[\sqrt{1-\lambda}\mathbb{G}_{1hj}(s,t) - \sqrt{\lambda}\mathbb{G}_{2hj}(s,t)\right]^2 \mathrm{d}\mu(t)\right\}^{1/2},$$

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}Q_{n_1,n_2,j} \rightsquigarrow \left\{\int_0^\tau \left[\sqrt{1-\lambda}\mathbb{G}_{1j}(t) - \sqrt{\lambda}\mathbb{G}_{2j}(t)\right]^2 \mathrm{d}\mu(t)\right\}^{1/2},$$

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}K_{n_1,n_2,hj}(s) \rightsquigarrow \sup_{t\in[s,\tau]}\left|\sqrt{1-\lambda}\mathbb{G}_{1hj}(s,t) - \sqrt{\lambda}\mathbb{G}_{2hj}(s,t)\right|,$$

and

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}K_{n_1,n_2,j} \rightsquigarrow \sup_{t\in[s,\tau]}\left|\sqrt{1-\lambda}\mathbb{G}_{1j}(t) - \sqrt{\lambda}\mathbb{G}_{2j}(t)\right|.$$

The asymptotic null distributions of the test statistics $Q'_{n_1,n_2,hj}(s)$, $Q'_{n_1,n_2,j}$, $K'_{n_1,n_2,hj}(s)$, and $K'_{n_1,n_2,j}$ are similar to those listed above. These asymptotic null distributions are quite intractable and of limited usefulness for hypothesis testing in practice. However, Theorem 2 and the continuous mapping theorem provide a way to generate realizations from these asymptotic null distributions. Calculation of the p value based on the statistics $Q_{n_1,n_2,hj}(s)$ and $K_{n_1,n_2,hj}(s)$ can be achieved via the multiplier process $\hat{C}_{n_1,n_2,hj}(s,t)$ using the following algorithm.

Algorithm 1. Choose a large integer B (say $B = 1000$) and for each $b = 1, \ldots, B$ repeat the steps

Step 1. Simulate sets of independent standard normal variables $\{\xi_{ip}^{(b)}\}_{i=1}^n$, $p = 1, 2$.

Step 2. Based on $\{\xi_{ip}^{(b)}\}_{i=1}^n$, $p = 1, 2$, calculate a realization

$$\hat{C}^{(b)}_{n_1,n_2,hj}(s,t) = \sqrt{\frac{n_2}{n_1+n_2}}\hat{W}_{hj}(t)\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{\gamma}_{i,1hj}(s,t)\xi^{(b)}_{i1} - \sqrt{\frac{n_1}{n_1+n_2}}\hat{W}_{hj}(t)\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\hat{\gamma}_{i,2hj}(s,t)\xi^{(b)}_{i2}, \quad t\in[s,\tau].$$

Once this process is complete, the p value can be approximated, depending on the type of test, as either

$$\frac{1}{B}\sum_{b=1}^{B} I\left(\left\{\int_s^\tau \left[\hat{C}^{(b)}_{n_1,n_2,hj}(s,t)\right]^2 \mathrm{d}\mu(t)\right\}^{1/2} \geq \sqrt{\frac{n_1n_2}{n_1+n_2}}Q_{n_1,n_2,hj}(s)\right),$$

or

$$\frac{1}{B}\sum_{b=1}^{B} I\left[\sup_{t\in[s,\tau]}\left|\hat{C}^{(b)}_{n_1,n_2,hj}(s,t)\right| \geq \sqrt{\frac{n_1n_2}{n_1+n_2}}K_{n_1,n_2,hj}(s)\right].$$

Alternatively, one can use the easier to implement cluster bootstrap using the following algorithm.

Algorithm 2. Choose a large integer B (say $B = 1000$) and for each $b = 1,\ldots,B$ repeat the steps

Step 1. Generate a cluster bootstrap estimated difference $\hat{\Delta}^{*(b)}_{n_1,n_2,hj}(s,t)$.

Step 2. Based on $\hat{\Delta}^{*(b)}_{n_1,n_2,hj}(s,t)$, calculate the cluster bootstrap

$$\hat{W}_{hj}(t)\left[\hat{\Delta}^{*(b)}_{n_1,n_2,hj}(s,t) - \hat{\Delta}_{n_1,n_2,hj}(s,t)\right], \quad t\in[s,\tau].$$

Once this process is complete, the p value can be approximated, depending on the type of test, as either

$$\frac{1}{B}\sum_{b=1}^{B} I\left[\left(\int_s^\tau \left\{\hat{W}_{hj}(t)\left[\hat{\Delta}^{*(b)}_{n_1,n_2,hj}(s,t) - \hat{\Delta}_{n_1,n_2,hj}(s,t)\right]\right\}^2 \mathrm{d}\mu(t)\right)^{1/2} \geq Q_{n_1,n_2,hj}(s)\right],$$

or

$$\frac{1}{B}\sum_{b=1}^{B} I\left\{\sup_{t\in[s,\tau]}\left|\hat{W}_{hj}(t)\left[\hat{\Delta}^{*(b)}_{n_1,n_2,hj}(s,t) - \hat{\Delta}_{n_1,n_2,hj}(s,t)\right]\right| \geq K_{n_1,n_2,hj}(s)\right\}.$$

Similar algorithms can be used for the remaining test statistics.

The L^2-norm-based and KS-type tests are consistent against any fixed alternative hypothesis. This statement is a consequence of Theorem 2, the uniform consistency of the population-averaged transition probability and state occupation probability estimators (Bakoyannis, 2021), the continuity of these tests in the differences

$\hat{\Delta}_{n_1,n_2,hj}(s,t)$, $\hat{\Delta}_{n_1,n_2,j}(t)$, $\hat{\Delta}'_{n_1,n_2,hj}(s,t)$, and $\hat{\Delta}'_{n_1,n_2,j}(t)$, and Lemma 14.15 in van der Vaart (2000).

3.2 Dependent groups

In many situations, the two groups under comparison are not independent. Such a situation arises in a multicenter randomized controlled trial, where, within each cluster (e.g., clinic), some cluster members receive the intervention of interest, and the remaining cluster members may receive placebo. For the situation with dependent groups, we have that $M_{i1} + M_{i2} = M_i$, $i = 1, \ldots, n$. In this case, there are two possibilities; (i) all clusters include processes from both groups, i.e., $M_{i1} \wedge M_{i2} > 0$ a.s., $i = 1, \ldots, n$, (complete cluster structure) and (ii) a subset of clusters include processes from one group only (incomplete cluster structure). We remind the reader that Bakoyannis (2021) only proposed a KS-type test for the situation with dependent groups with complete cluster structure.

3.2.1 Complete cluster structure

In this subsection, we assume that $M_{i1} \wedge M_{i2} > 0$ almost surely. Based on this study setup, define the estimators of the pointwise between-group difference with respect to the population-averaged transition probabilities as

$$\hat{\Delta}_{n,hj}(s,t) = \left[\hat{P}_{n,1hj}(s,t) - \hat{P}_{n,2hj}(s,t)\right], \quad t \in [s,\tau],$$

where $\hat{P}_{n,phj}$, $p = 1, 2$, is the estimator of $P_{0,phj}$ from the pth group and

$$\hat{\Delta}'_{n,hj}(s,t) = \left[\hat{P}'_{n,1hj}(s,t) - \hat{P}'_{n,2hj}(s,t)\right], \quad t \in [s,\tau],$$

where $\hat{P}'_{n,phj}$, $p = 1, 2$, is the estimator of $P'_{0,phj}$ from the pth group, for some $s \in [0,\tau)$. Similarly, define the differences between the population-averaged state occupation probabilities as

$$\hat{\Delta}_{n,j}(t) = \left[\hat{P}_{n,1j}(t) - \hat{P}_{n,2j}(t)\right], \quad t \in [0,\tau],$$

where $\hat{P}_{n,pj}$, $p = 1, 2$, is the estimator of $P_{0,pj}$ from the pth group, and

$$\hat{\Delta}'_{n,j}(t) = \left[\hat{P}'_{n,1j}(t) - \hat{P}'_{n,2j}(t)\right], \quad t \in [0,\tau],$$

where $\hat{P}'_{n,pj}$, $p = 1, 2$, is the estimator of $P'_{0,pj}$ from the pth group. The corresponding nonparametric cluster bootstrap realizations of the above differences are denoted by $\hat{\Delta}^*_{n,hj}(s,t)$, $\hat{\Delta}'^*_{n,hj}(s,t)$, $\hat{\Delta}^*_{n,j}(t)$, and $\hat{\Delta}'^*_{n,j}(t)$. It is important to note that these nonparametric cluster bootstrap realizations are generated by randomly sampling n clusters with replacement. Based on these differences, we define the following linear test statistics:

$$Z_{n,hj}(s) = \int_s^\tau \hat{W}_{hj}(t)\hat{\Delta}_{n,hj}(s,t)\mathrm{d}\mu(t), \quad \text{for some } s \in [0,\tau),$$
$$Z'_{n,hj}(s) = \int_s^\tau \hat{W}'_{hj}(t)\hat{\Delta}'_{n,hj}(s,t)\mathrm{d}\mu(t), \quad \text{for some } s \in [0,\tau),$$
$$Z_{n,j} = \int_s^\tau \hat{W}_j(t)\hat{\Delta}_{n,j}(t)\mathrm{d}\mu(t),$$

and

$$Z_{n,j} = \int_s^\tau \hat{W}'_j(t)\hat{\Delta}'_{n,j}(t)\mathrm{d}\mu(t),$$

where the weights are defined as in Sect. 3.1. Theorem 3 states the asymptotic null distributions of these linear tests for the dependent-groups case with complete cluster structure.

Theorem 3 *Suppose that conditions C1–C7 hold. Then, under the null hypothesis and for any* $h \in \mathcal{T}^e$, $j \in \mathcal{S}$, *and some* $s \in [0,\tau)$, *the following hold*

(i) $\sqrt{n}Z_{n,hj}(s) \rightsquigarrow Z_{hj}(s)$ *as* $n \to \infty$, *where* $Z_{hj}(s) \sim N(0, \eta^2_{hj}(s))$ *and*

$$\eta^2_{hj}(s) = E\left\{\int_s^\tau W_{hj}(t)[\gamma_{1,1hj}(s,t) - \gamma_{1,2hj}(s,t)]\mathrm{d}\mu(t)\right\}^2.$$

(ii) $\sqrt{n}Z_{n,j} \rightsquigarrow Z_j$ *as* $n \to \infty$, *where* $Z_j \sim N(0, \eta^2_j)$ *and*

$$\eta^2_j = E\left\{\int_0^\tau W_j(t)[\psi_{1,1j}(t) - \psi_{1,2j}(t)]\mathrm{d}\mu(t)\right\}^2.$$

The proof of Theorem 3 is given in the Supplementary Material. Consistent (in probability) estimators of the asymptotic variances in Theorem 3 are

$$\hat{\eta}^2_{hj}(s) = \frac{1}{n}\sum_{i=1}^n \left\{\int_s^\tau \hat{W}_{hj}(t)[\hat{\gamma}_{i,1hj}(s,t) - \hat{\gamma}_{i,2hj}(s,t)]\mathrm{d}\mu(t)\right\}^2,$$

and

$$\hat{\eta}^2_j = \frac{1}{n}\sum_{i=1}^n \left\{\int_0^\tau \hat{W}_j(t)[\hat{\psi}_{i,1j}(t) - \hat{\psi}_{i,2j}(t)]\mathrm{d}\mu(t)\right\}^2.$$

Alternatively, by Theorem 2 in Bakoyannis (2021) and the bootstrap continuous mapping theorem (Theorem 10.8 in Kosorok, 2008), these variances can be estimated as sample variances based on a number of nonparametric cluster bootstrap realizations $\int_s^\tau \hat{W}_{hj}(t)[\hat{\Delta}^*_{n,hj}(s,t) - \hat{\Delta}_{n,hj}(s,t)]\mathrm{d}\mu(t)$ and $\int_0^\tau \hat{W}_j(t)[\hat{\Delta}^*_{n,j}(t) - \hat{\Delta}_{n,j}(t)]\mathrm{d}\mu(t)$,

respectively. Based on any of these variance estimators and Theorem 3, it is easy to construct an asymptotic Z-test for testing the null hypothesis of interest as usual. Using the same arguments as those used in the proof of Theorem 3, it can also be shown that a similar version of this theorem holds for the test statistics $Z'_{n,hj}(s)$ and $Z'_{n,j}$.

As for the case with independent groups, we define the L^2-norm-based tests

$$Q_{n,hj}(s) = \left\{ \int_s^\tau \left[\hat{W}_{hj}(t) \hat{\Delta}_{n,hj}(s,t) \right]^2 \mathrm{d}\mu(t) \right\}^{1/2}, \quad \text{for some } s \in [0,\tau),$$

and

$$Q_{n,j} = \left\{ \int_0^\tau \left[\hat{W}_j(t) \hat{\Delta}_{n,j}(t) \right]^2 \mathrm{d}\mu(t) \right\}^{1/2}.$$

The KS-type tests by Bakoyannis (2021) are $K_{n,hj}(s) = \sup_{t\in[s,\tau]} |\hat{W}_{hj}(t)\hat{\Delta}_{n,hj}(s,t)|$, for some $s \in [0,\tau)$, and $K_{n,j} = \sup_{t\in[0,\tau]} |\hat{W}_j(t)\hat{\Delta}_{n,j}(t)|$. The corresponding tests for $\hat{\Delta}'_{n,hj}(s,t)$ and $\hat{\Delta}'_{n,j}(t)$, denoted by $Q'_{n,hj}(s)$, $K'_{n,hj}(s)$, and $Q'_{n,j}$, $K'_{n,j}$, are defined in a similar manner. Conducting hypothesis testing with these tests can be based on a resampling scheme that utilizes the estimated multiplier processes

$$\hat{C}_{n,hj}(s,t) = \hat{W}_{hj}(t) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\hat{\gamma}_{i,1hj}(s,t) - \hat{\gamma}_{i,2hj}(s,t)] \xi_i \quad t \in [s,\tau],$$

for some $s \in [0,\tau)$, where ξ_i, $i = 1,\ldots,n$, are independent standard normal variables, and

$$\hat{C}_{n,j}(t) = \hat{W}_j(t) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\hat{\psi}_{i,1j}(t) - \hat{\psi}_{i,2j}(t)] \xi_i \quad t \in [0,\tau].$$

Similarly, one can define the estimated processes $\hat{C}_{n,hj}(s,t)$ and $\hat{C}'_{n,j}(t)$ which correspond to the tests for $\hat{\Delta}'_{n,hj}(s,t)$ and $\hat{\Delta}'_{n,j}(t)$. Under conditions C1–C7 and by Theorem 3 in Bakoyannis (2021), these processes converge weakly, conditionally on the observed data, to the null limiting processes of the corresponding test statistics. This fact along with the continuous mapping theorem can be used for the calculation of p values via similar algorithms to those described in Sect. 3.1. Finally, using the same arguments to those presented in the end of Sect. 3.1, the L^2-norm-based and KS-type tests are consistent for any fixed alternative hypothesis.

3.2.2 Incomplete cluster structure

Here, we relax the assumption that $M_{i1} \wedge M_{i2} > 0$ almost surely, and allow some clusters to have observations from one group only. Suppose that n_1 clusters involve processes from group 1 only, n_2 clusters involve processes from group 2 only, and n clusters include processes from both groups. As in previous research on the simpler

location problem with incompletely paired observations (Fong et al., 2018), we assume that the unobserved groups in some clusters are missing completely at random (Little and Rubin, 2019). Under this setting, we propose three hybrid tests that utilize the tests for independent and dependent groups with complete cluster structure. The proposed hybrid test statistics for the hypothesis $H_0 : P_{0,1hj}(s,\cdot) = P_{0,2hj}(s,\cdot)$ are

$$\left[\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{Z_{n_1,n_2,hj}(s)}{\hat{\omega}_{hj}(s)}\right]^2 + \left[\sqrt{n}\frac{Z_{n,hj}(s)}{\hat{\eta}_{hj}(s)}\right]^2,$$
$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}Q_{n_1,n_2,hj}(s) + \sqrt{n}Q_{n,hj}(s),$$

and

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}K_{n_1,n_2,hj}(s) + \sqrt{n}K_{n,hj}(s).$$

The hybrid tests for the other null hypotheses have a similar structure. Theorem 4 provides the basis for conducting hypothesis testing based on the above tests.

Theorem 4 *Suppose that conditions C1–C7 hold and that* $n_1/(n_1+n_2) \to \lambda \in (0,1)$ *as* $n_1 \wedge n_2 \to \infty$. *Then, under the null hypothesis and for any* $h \in \mathcal{T}^\tau$, $j \in \mathcal{S}$, *and some* $s \in [0,\tau)$, *it follows that as* $n_1 \wedge n_2 \wedge n \to \infty$

(i) $\left[\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{Z_{n_1,n_2,hj}(s)}{\hat{\omega}_{hj}(s)}\right]^2 + \left[\sqrt{n}\frac{Z_{n,hj}(s)}{\hat{\eta}_{hj}(s)}\right]^2 \rightsquigarrow \chi^2_2.$

(ii) *Both random sequences* $\{\int_s^\tau [\hat{C}_{n_1,n_2,hj}(s,t)]^2 \mathrm{d}\mu(t)\}^{1/2} + \{\int_s^\tau [\hat{C}_{n,hj}(s,t)]^2 \mathrm{d}\mu(t)\}^{1/2}$ *and*

$$\left(\int_s^\tau \left\{\hat{W}_{hj}(t)\left[\hat{\Delta}^*_{n_1,n_2,hj}(s,t) - \hat{\Delta}_{n_1,n_2,hj}(s,t)\right]\right\}^2 \mathrm{d}\mu(t)\right)^{1/2}$$
$$+ \left(\int_s^\tau \left\{\hat{W}_{hj}(t)\left[\hat{\Delta}^*_{n,hj}(s,t) - \hat{\Delta}_{n,hj}(s,t)\right]\right\}^2 \mathrm{d}\mu(t)\right)^{1/2}$$

converge weakly, conditionally on the observed data, to the asymptotic null distribution of $\sqrt{\frac{n_1 n_2}{n_1+n_2}}Q_{n_1,n_2,hj}(s) + \sqrt{n}Q_{n,hj}(s)$.

(iii) *Both random sequences* $\sup_{t\in[s,\tau]} |\hat{C}_{n_1,n_2,hj}(s,t)| + \sup_{t\in[s,\tau]} |\hat{C}_{n,hj}(s,t)|$ *and*

$$\sup_{t\in[s,\tau]}\left|\hat{W}_{hj}(t)\left[\hat{\Delta}^*_{n_1,n_2,hj}(s,t) - \hat{\Delta}_{n_1,n_2,hj}(s,t)\right]\right|$$
$$+ \sup_{t\in[s,\tau]}\left|\hat{W}_{hj}(t)\left[\hat{\Delta}^*_{n,hj}(s,t) - \hat{\Delta}_{n,hj}(s,t)\right]\right|$$

converge weakly, conditionally on the observed data, to the asymptotic null distribution of $\sqrt{\frac{n_1 n_2}{n_1+n_2}}K_{n_1,n_2,hj}(s) + \sqrt{n}K_{n,hj}(s)$.

Theorem 4 follows from Theorems 1–3, Theorem 3 in Bakoyannis (2021), the continuous mapping theorem, the bootstrap continuous mapping theorem (Kosorok, 2008), and the assumption of independence across clusters. In light of Theorem 4, it is easy to conduct hypothesis testing based on the hybrid tests for situations with incomplete cluster structure. Calculation of the p value based on the multiplier processes and the cluster bootstraps in parts (ii) and (iii) of the latter theorem can be performed using similar algorithms to those provided in Sect. 3.1.

4 Simulation studies

A series of simulation experiments was conducted to evaluate the small-sample performance of the proposed tests and compare them with the KS-type test for dependent groups with complete cluster structure by Bakoyannis (2021). The simulation setup was similar to that in Bakoyannis (2021). Specifically, we considered a non-Markov illness-death model with state space $\mathcal{S} = \{1, 2, 3\}$ and absorbing state $\mathcal{T} = \{3\}$, in a study with clustered observations and ICS. In this simulation study, we choose $n = 20, 40, 80$ clusters, which are considered as small, or relatively small numbers of clusters. The cluster sizes M_i, $i = 1, \ldots, n$, were simulated from the discrete uniform distributions $\mathcal{U}(5, 15)$ or $\mathcal{U}(10, 30)$, producing scenarios with 5 to 15 or 10 to 30 observations per cluster, respectively. To induce within-cluster dependence and simulate non-Markov processes, we generated cluster-specific frailties v_i, $i = 1, \ldots, n$, from the Gamma distribution with shape and scale parameters equal

Table 1 Simulation results for two independent groups (scenario 1) regarding the empirical type I error of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,112}(0.5, \cdot) = P_{0,212}(0.5, \cdot)$ and $H_0 : P'_{0,112}(0.5, \cdot) = P'_{0,212}(0.5, \cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

$n_1 = n_2$	F_M	Method	$P_{0,p12}(0.5, \cdot)$, $p = 1, 2$			$P'_{0,p12}(0.5, \cdot)$, $p = 1, 2$		
			Linear	L^2	KS	Linear	L^2	KS
20	$\mathcal{U}[5, 15]$	IF	0.055	0.055	0.044	0.057	0.054	0.044
		CB	0.049	0.048	0.033	0.054	0.051	0.040
	$\mathcal{U}[10, 30]$	IF	0.063	0.067	0.055	0.071	0.068	0.055
		CB	0.062	0.056	0.051	0.069	0.061	0.049
40	$\mathcal{U}[5, 15]$	IF	0.051	0.053	0.044	0.048	0.044	0.041
		CB	0.049	0.052	0.042	0.049	0.043	0.035
	$\mathcal{U}[10, 30]$	IF	0.051	0.043	0.050	0.049	0.038	0.051
		CB	0.050	0.042	0.042	0.049	0.038	0.047
80	$\mathcal{U}[5, 15]$	IF	0.047	0.049	0.043	0.052	0.053	0.049
		CB	0.046	0.046	0.038	0.054	0.053	0.046
	$\mathcal{U}[10, 30]$	IF	0.049	0.053	0.047	0.048	0.047	0.045
		CB	0.052	0.054	0.047	0.045	0.049	0.041

n: Number of clusters, F_M: Distribution of the cluster size

Table 2 Simulation results for two independent groups (scenario 1) regarding the empirical power of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,112}(0.5, \cdot) = P_{0,212}(0.5, \cdot)$ and $H_0 : P'_{0,112}(0.5, \cdot) = P'_{0,212}(0.5, \cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

$n_1 = n_2$	F_M	Method	$P_{0,p12}(0.5, \cdot)$, $p = 1, 2$			$P'_{0,p12}(0.5, \cdot)$, $p = 1, 2$		
			Linear	L^2	KS	Linear	L^2	KS
20	$\mathcal{U}[5, 15]$	IF	0.261	0.218	0.156	0.257	0.214	0.144
		CB	0.251	0.216	0.136	0.252	0.209	0.138
	$\mathcal{U}[10, 30]$	IF	0.368	0.327	0.254	0.346	0.297	0.222
		CB	0.360	0.310	0.216	0.337	0.290	0.201
40	$\mathcal{U}[5, 15]$	IF	0.460	0.406	0.310	0.444	0.380	0.278
		CB	0.457	0.397	0.303	0.441	0.373	0.265
	$\mathcal{U}[10, 30]$	IF	0.648	0.612	0.504	0.625	0.589	0.477
		CB	0.638	0.603	0.476	0.619	0.576	0.464
80	$\mathcal{U}[5, 15]$	IF	0.747	0.704	0.577	0.703	0.658	0.530
		CB	0.743	0.699	0.559	0.700	0.659	0.522
	$\mathcal{U}[10, 30]$	IF	0.898	0.875	0.791	0.875	0.848	0.759
		CB	0.897	0.878	0.791	0.878	0.852	0.761

n: Number of clusters; F_M: Distribution of the cluster size

Table 3 Simulation results for two independent groups (scenario 1) regarding the empirical type I error of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,12}(\cdot) = P_{0,22}(\cdot)$ and $H_0 : P'_{0,12}(\cdot) = P'_{0,22}(\cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

$n_1 = n_2$	F_M	Method	$P_{0,p2}(\cdot)$, $p = 1, 2$			$P'_{0,p2}(\cdot)$, $p = 1, 2$		
			Linear	L^2	KS	Linear	L^2	KS
20	$\mathcal{U}[5, 15]$	IF	0.060	0.062	0.055	0.065	0.066	0.052
		CB	0.061	0.058	0.051	0.067	0.066	0.051
	$\mathcal{U}[10, 30]$	IF	0.061	0.069	0.054	0.065	0.061	0.054
		CB	0.064	0.066	0.050	0.066	0.057	0.048
40	$\mathcal{U}[5, 15]$	IF	0.044	0.048	0.048	0.040	0.036	0.038
		CB	0.043	0.045	0.048	0.038	0.036	0.035
	$\mathcal{U}[10, 30]$	IF	0.056	0.053	0.046	0.052	0.048	0.047
		CB	0.056	0.052	0.048	0.051	0.050	0.048
80	$\mathcal{U}[5, 15]$	IF	0.053	0.053	0.045	0.057	0.049	0.041
		CB	0.049	0.054	0.042	0.061	0.056	0.041
	$\mathcal{U}[10, 30]$	IF	0.055	0.056	0.055	0.044	0.049	0.048
		CB	0.053	0.053	0.050	0.044	0.048	0.047

n: Number of clusters; F_M: Distribution of the cluster size

Table 4 Simulation results for two independent groups (scenario 1) regarding the empirical power of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,12}(\cdot) = P_{0,22}(\cdot)$ and $H_0 : P'_{0,12}(\cdot) = P'_{0,22}(\cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

$n_1 = n_2$	F_M	Method	$P_{0,p2}(\cdot)$, $p = 1, 2$			$P'_{0,p2}(\cdot)$, $p = 1, 2$		
			Linear	L^2	KS	Linear	L^2	KS
20	$\mathcal{U}[5, 15]$	IF	0.526	0.494	0.400	0.517	0.476	0.374
		CB	0.524	0.492	0.385	0.518	0.471	0.363
	$\mathcal{U}[10, 30]$	IF	0.613	0.576	0.497	0.601	0.565	0.474
		CB	0.610	0.566	0.489	0.601	0.566	0.470
40	$\mathcal{U}[5, 15]$	IF	0.804	0.775	0.699	0.778	0.744	0.655
		CB	0.796	0.768	0.691	0.779	0.742	0.656
	$\mathcal{U}[10, 30]$	IF	0.900	0.880	0.826	0.890	0.871	0.812
		CB	0.902	0.877	0.818	0.892	0.868	0.805
80	$\mathcal{U}[5, 15]$	IF	0.969	0.966	0.935	0.964	0.961	0.914
		CB	0.970	0.967	0.935	0.964	0.959	0.910
	$\mathcal{U}[10, 30]$	IF	0.995	0.993	0.987	0.995	0.995	0.985
		CB	0.995	0.993	0.984	0.995	0.995	0.981

n: Number of clusters; F_M: Distribution of the cluster size

to 1, and simulated illness-death processes using the conditional (on the frailty) cumulative transition intensities $A_{0,12}(t;v_i) = [0.25 + 0.25 \times I\{m_i \leq E(M_1)\}]v_i t$, $A_{0,23}(t;v_i) = 0.5v_i t$, and $A_{0,13}(t;v_i) = 0.25v_i t$, $i = 1, \ldots, n$. The dependence of $A_{0,12}(t;v_i)$ on the cluster size induced scenarios with ICS. In addition, we simulated independent right censoring times from the uniform distribution $U(0, 3)$. Two main scenarios according to the study design were considered: 1) a cluster randomized trial where the two groups were independent and 2) a multicenter randomized trial where the two groups were dependent with a complete cluster structure. In both scenarios, we used a 1:1 group allocation ratio. In this simulation study, we focused on the between-group comparison of the population-averaged transition probabilities $P'_{0,12}(0.5, t)$ and $P_{0,12}(0.5, t)$, and the population-averaged state occupation probabilities $P_{0,2}(t)$ and $P'_{0,2}(t)$. Data under the alternative hypothesis were simulated using the cumulative transition intensity $A_{0,p12}(t;v_i) = [0.25 + 0.5 \times I(p = 2) + 0.25 \times I\{m_i \leq E(M_1)\}]v_i t$, $p = 1, 2$, which depends on treatment arm p. Estimation of the transition probabilities was performed using the landmark version of the proposed estimators as described in Sect. 2. For each scenario, we simulated 1000 datasets and, in each dataset, we tested the null hypothesis of interest with the proposed tests. The KS-type test by Bakoyannis (2021) is not applicable in scenario 1 with independent groups and, thus, was only considered in scenario 2 (dependent groups). The calculation of the p values from the linear tests was based on the corresponding asymptotic normal distribution under the null, where the variance was estimated by both the closed-form estimators that utilize the empirical versions of the influence functions and the nonparametric cluster bootstrap with 1000 replications. For the calculation of

Table 5 Simulation results for two dependent groups (scenario 2) regarding the empirical type I error of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,112}(0.5, \cdot) = P_{0,212}(0.5, \cdot)$ and $H_0 : P'_{0,112}(0.5, \cdot) = P'_{0,212}(0.5, \cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

n	F_M	Method	$P_{0,p12}(0.5, \cdot)$, $p = 1, 2$			$P'_{0,p12}(0.5, \cdot)$, $p = 1, 2$		
			Linear	L^2	KS*	Linear	L^2	KS*
20	$\mathcal{U}[5, 15]$	IF	0.050	0.049	0.046	0.055	0.048	0.040
		CB	0.045	0.042	0.041	0.055	0.042	0.038
	$\mathcal{U}[10, 30]$	IF	0.066	0.056	0.042	0.071	0.055	0.046
		CB	0.060	0.048	0.037	0.068	0.052	0.043
40	$\mathcal{U}[5, 15]$	IF	0.051	0.057	0.052	0.048	0.047	0.043
		CB	0.049	0.052	0.037	0.050	0.046	0.040
	$\mathcal{U}[10, 30]$	IF	0.041	0.036	0.040	0.046	0.037	0.042
		CB	0.038	0.038	0.037	0.045	0.040	0.038
80	$\mathcal{U}[5, 15]$	IF	0.046	0.040	0.048	0.046	0.040	0.046
		CB	0.047	0.041	0.049	0.042	0.038	0.044
	$\mathcal{U}[10, 30]$	IF	0.050	0.055	0.059	0.054	0.057	0.052
		CB	0.051	0.054	0.061	0.057	0.054	0.050

n: Number of clusters; F_M: Distribution of the cluster size

*Kolmogorov–Smirnov-type test by Bakoyannis (2021)

Table 6 Simulation results for two dependent groups (scenario 2) regarding the empirical power of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,112}(0.5, \cdot) = P_{0,212}(0.5, \cdot)$ and $H_0 : P'_{0,112}(0.5, \cdot) = P'_{0,212}(0.5, \cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

n	F_M	Method	$P_{0,p12}(0.5, \cdot)$, $p = 1, 2$			$P'_{0,p12}(0.5, \cdot)$, $p = 1, 2$		
			Linear	L^2	KS*	Linear	L^2	KS*
20	$\mathcal{U}[5, 15]$	IF	0.202	0.169	0.108	0.205	0.161	0.093
		CB	0.198	0.158	0.092	0.197	0.150	0.089
	$\mathcal{U}[10, 30]$	IF	0.406	0.335	0.218	0.349	0.290	0.193
		CB	0.395	0.311	0.188	0.345	0.268	0.164
40	$\mathcal{U}[5, 15]$	IF	0.391	0.327	0.233	0.340	0.287	0.214
		CB	0.388	0.308	0.221	0.339	0.272	0.206
	$\mathcal{U}[10, 30]$	IF	0.626	0.553	0.410	0.580	0.523	0.364
		CB	0.622	0.544	0.386	0.577	0.508	0.346
80	$\mathcal{U}[5, 15]$	IF	0.660	0.606	0.428	0.598	0.535	0.356
		CB	0.658	0.599	0.414	0.589	0.530	0.351
	$\mathcal{U}[10, 30]$	IF	0.913	0.868	0.723	0.867	0.823	0.644
		CB	0.911	0.868	0.713	0.862	0.815	0.639

n: number of clusters; F_M: distribution of the cluster size

*Kolmogorov–Smirnov-type test by Bakoyannis (2021)

Table 7 Simulation results for two dependent groups (scenario 2) regarding the empirical type I error of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,12}(\cdot) = P_{0,22}(\cdot)$ and $H_0 : P'_{0,12}(\cdot) = P'_{0,22}(\cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

n	F_M	Method	$P_{0,p2}(\cdot)$, $p = 1, 2$			$P'_{0,p2}(\cdot)$, $p = 1, 2$		
			Linear	L^2	KS*	Linear	L^2	KS*
20	$\mathcal{U}[5, 15]$	IF	0.069	0.063	0.045	0.060	0.051	0.049
		CB	0.068	0.054	0.042	0.061	0.051	0.050
	$\mathcal{U}[10, 30]$	IF	0.063	0.052	0.040	0.067	0.051	0.044
		CB	0.061	0.053	0.039	0.065	0.047	0.040
40	$\mathcal{U}[5, 15]$	IF	0.058	0.055	0.044	0.056	0.045	0.037
		CB	0.057	0.057	0.041	0.056	0.046	0.039
	$\mathcal{U}[10, 30]$	IF	0.061	0.056	0.048	0.059	0.055	0.044
		CB	0.060	0.050	0.046	0.057	0.053	0.046
80	$\mathcal{U}[5, 15]$	IF	0.042	0.051	0.048	0.049	0.047	0.049
		CB	0.040	0.050	0.046	0.050	0.047	0.047
	$\mathcal{U}[10, 30]$	IF	0.057	0.055	0.053	0.059	0.058	0.059
		CB	0.056	0.055	0.053	0.060	0.061	0.055

n: number of clusters; F_M: distribution of the cluster size

*Kolmogorov–Smirnov-type test by Bakoyannis (2021)

Table 8 Simulation results for two dependent groups (scenario 2) regarding the empirical power of the proposed linear test (Linear), L^2-norm-based test (L^2), and Kolmogorov–Smirnov-type test (KS) for $H_0 : P_{0,12}(\cdot) = P_{0,22}(\cdot)$ and $H_0 : P'_{0,12}(\cdot) = P'_{0,22}(\cdot)$ at the $\alpha = 0.05$ level. Significance levels were calculated based on either the empirical influence functions (IF) or the nonparametric cluster bootstrap (CB)

n	F_M	Method	$P_{0,p2}(\cdot)$, $p = 1, 2$			$P'_{0,p2}(\cdot)$, $p = 1, 2$		
			Linear	L^2	KS*	Linear	L^2	KS*
20	$\mathcal{U}[5, 15]$	IF	0.489	0.449	0.352	0.464	0.430	0.331
		CB	0.486	0.445	0.339	0.462	0.433	0.337
	$\mathcal{U}[10, 30]$	IF	0.791	0.737	0.634	0.748	0.714	0.598
		CB	0.781	0.743	0.625	0.744	0.719	0.601
40	$\mathcal{U}[5, 15]$	IF	0.809	0.771	0.666	0.755	0.719	0.612
		CB	0.809	0.773	0.659	0.750	0.709	0.603
	$\mathcal{U}[10, 30]$	IF	0.971	0.962	0.905	0.956	0.931	0.874
		CB	0.970	0.958	0.906	0.955	0.927	0.873
80	$\mathcal{U}[5, 15]$	IF	0.973	0.965	0.916	0.949	0.934	0.870
		CB	0.972	0.965	0.917	0.949	0.933	0.864
	$\mathcal{U}[10, 30]$	IF	1.000	1.000	0.995	1.000	0.998	0.991
		CB	1.000	0.999	0.994	0.999	0.996	0.990

n: Number of clusters; F_M: Distribution of the cluster size

*Kolmogorov–Smirnov-type test by Bakoyannis (2021)

p values from the L^2-norm-based and the KS-type tests, we used both the multiplier processes that depend on the empirical influence functions, with 1000 simulated sets $\{\xi_i\}_{i=1}^n$ of standard normal variables, and the cluster bootstrap with 1000 bootstrap replications.

Simulation results under scenario 1 (independent groups) are presented in Tables 1, 2, 3 and 4. The empirical type I error rates of the tests were close to the 0.05 level in all cases for both transition (Table 1) and state occupation probabilities (Table 3). This indicates that the approximation of the null distributions of the tests by their corresponding asymptotic distributions was particularly good, even in cases with a small number of clusters. As expected, the empirical power (Tables 2, 4) increased with sample size. The linear and the L^2-norm-based tests exhibited a substantially larger power compared to the KS-type test. The linear test was also somewhat more powerful compared to the L^2-norm-based test, particularly in cases with smaller sample sizes. In addition, the weighted by cluster size tests exhibited slightly lower power levels compared to their unweighted counterparts. This is attributed to the additional variability of the cluster sizes in the weights.

Simulation results under scenario 2 (dependent groups) are summarized in Tables 5, 6, 7 and 8. The results from scenario 2 were similar to those from scenario 1. The empirical type I error rates (Tables 5, 7) were close to the nominal level, even in cases with a small number of clusters and the empirical power levels (Tables 6, 8) increased with sample size. The linear and the L^2-norm-based tests were substantially more powerful compared to the KS-type test by Bakoyannis (2021). Furthermore, the linear test was somewhat more powerful compared to the L^2-norm-based test.

In summary, our simulation experiments provide numerical evidence that the proposed tests work well even with a small number of clusters and under ICS and non-Markov processes. In addition, our tests are substantially more powerful compared to the KS-type test by Bakoyannis (2021).

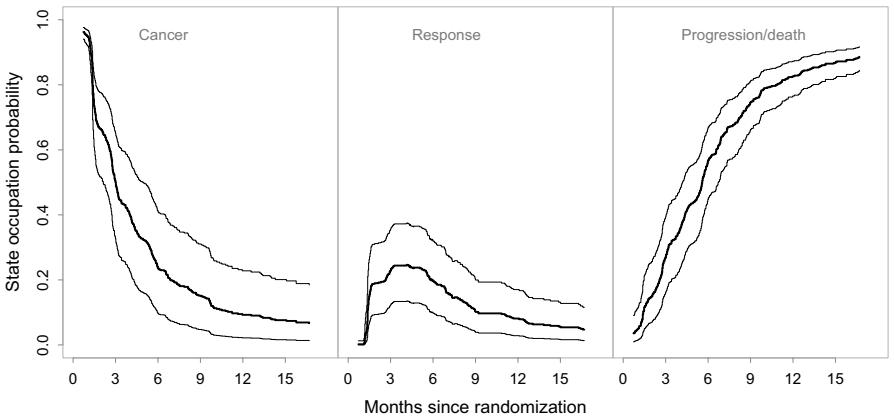

Fig. 1 Multicenter SPECTRUM study: Overall population-averaged state occupation probabilities, with the 95% simultaneous confidence bands

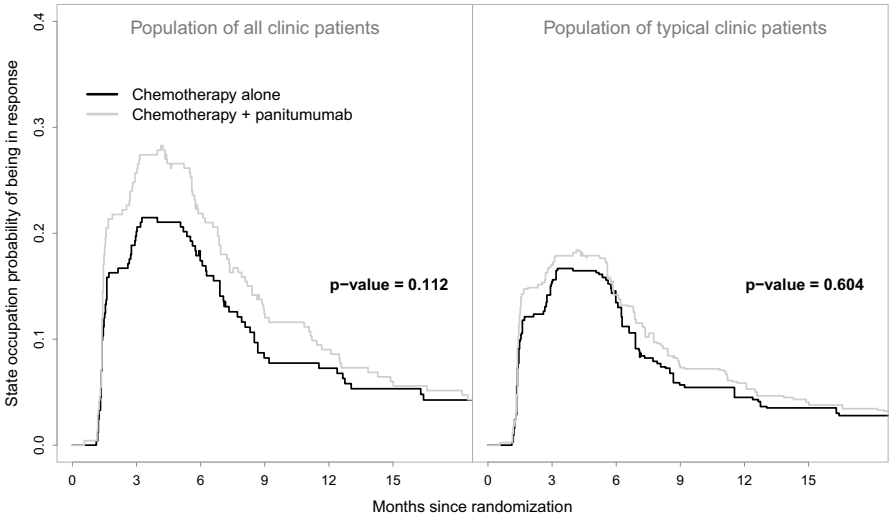

Fig. 2 Population-averaged state occupation probabilities of tumor response by treatment group in the multicenter SPECTRUM study, along with the p value from the linear test for dependent groups

5 Data application

The proposed tests were applied to the data from the multicenter SPECTRUM trial (Vermorken et al., 2013), an open-label Phase III randomized trial, conducted to evaluate the efficacy of the combination of chemotherapy with panitumumab, compared to chemotherapy alone, in terms of the probability of tumor response in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck. In the subset of the data which was available to us, there were 72 clinics and 479 patients. Of these patients, 243 were in the chemotherapy group and 236 in the chemotherapy plus panitumumab group. These groups were dependent with a complete cluster structure, that is each of the 72 clinics involved patients from both groups. Throughout the follow-up period, 126 patients experienced response at some point, 422 patients experienced a disease progression or died, and 57 patients were right-censored. No left truncation was present in this dataset. The data were analyzed under the illness-death model, with tumor response being the transient state of interest and progression or death being the absorbing state. The estimates of the population-averaged state occupation probabilities over the ACM population (i.e., the population of all clinic patients) from the full sample are depicted in Fig. 1. These estimates illustrate the history of disease under treatment. The population-averaged probabilities of tumor response by treatment group, both for the ACM and the TCM (i.e., typical clinic patients) populations, are depicted in Fig. 2. In both populations, the population-averaged probability of being in tumor response appears to be higher in the chemotherapy plus panitumumab group. Also, Fig. 2 provides some evidence for ICS in this dataset, as the state occupation probabilities for the two populations appear to be different. More precisely, since larger clinics tend to dominate the ACM population, larger clinics appear to have a higher probability of response in general.

In addition, the difference between the two groups appears to be more pronounced in larger clinics. According to the linear test for dependent groups, the more pronounced difference in the ACM population is not statistically significant (p value = 0.112). The corresponding p values for the L^2-norm-based and the KS-type test for dependent data were 0.114 and 0.176, respectively. The p value from the Kolmogorov–Smirnov-type test was larger and this is in accordance with the results from our simulation experiments. As expected, the linear test for the TCM population, where the difference between the two groups appears less pronounced, does not provide a statistically significant result (p value = 0.604).

6 Discussion

In this work, we addressed the issue of nonparametric two-sample testing for population-averaged transition and state occupation probabilities for multistate processes with clustered, right-censored, and/or left-truncated data. We proposed tests for situations with both independent and dependent groups, with and without complete cluster structure. For each case, we proposed a linear test, an L^2-norm-based test, and a KS-type test. The proposed tests do not impose assumptions regarding the structure of the within-cluster dependence, and are applicable under ICS, and for both Markov and non-Markov processes. These characteristics are crucial in many applications, such as the SPECTRUM trial analyzed in Sect. 5. The asymptotic null distributions of the tests were established using empirical process theory. Rigorous procedures for the calculation of p values were proposed, and the L^2-norm-based and KS-type tests were argued to be consistent against any fixed alternative hypothesis. Simulation experiments under complex settings showed that the proposed tests work well, even under a small number of clusters. In addition, even though the linear tests may not be consistent against alternatives with crossing transition and state occupation probability functions, they were shown to be substantially more powerful compared to the KS-type tests under alternatives with non-crossing probabilities. The tests were illustrated using a motivating dataset from a multicenter randomized controlled trial.

The nonparametric literature on multistate processes with independent data is rich (Aalen and Johansen, 1978; Glidden, 2002; Tattar and Vaman, 2014; Bluhmki et al., 2018, 2019; Bakoyannis, 2020). However, to the best of our knowledge, only Bakoyannis (2021) has proposed a two-sample nonparametric procedure for multistate processes with clustered data. Nevertheless, this test has two important limitations. First, by virtue of being a KS-type test, it may not be the most powerful test for situations with non-crossing transition and state occupation probability functions. Second, this test is only applicable to situations with dependent groups with complete cluster structure. It is not applicable for problems with independent groups, such as cluster randomized trials, or with dependent groups with incomplete cluster structure. In this work, we have addressed all these limitations. We proposed a linear test and an L^2-norm-based test that can be substantially more powerful compared to the KS-type test by Bakoyannis (2021) in settings with non-crossing probability functions, as shown in our simulation experiments. Furthermore, we addressed,

for the first time, the issue of nonparametric two-sample comparison for clustered multistate processes, where the two groups under comparison are either independent or dependent with incomplete cluster structure. From an applied standpoint, the linear test may be preferable over the L^2-norm-based and KS-type tests, because a statistically significant difference based on the former implies that one group spends more time in a particular state. This is not necessarily true for the L^2-norm-based and KS-type tests.

In this article, we assumed that right censoring is independent of the multistate process of interest and the cluster size, which may be violated in practice. To address this, our work can be extended by incorporating inverse probability of censoring weighting techniques to account for dependent censoring (Datta and Satten, 2002). In this case, the influence functions of the test statistics will involve the influence functions of the chosen estimator for the model of the censoring distribution. There is a number of additional practically important issues that were not addressed in this work. First, statistical study design issues remain, such as sample size calculation for trials with clustered multistate processes. Second, many trials involve stratified randomization or minimization and this needs to be taken into account into the testing procedure (Kahan and Morris, 2012). Third, there are trials that involve more than two interventions. Fourth, there may be an association between different clusters, as for example between two clinics in close proximity. Finally, multistate event processes may often depend on time-dependent covariates (Studer et al., 2018), and covariate-dependent testing can be crucial there. These issues require further methodology development, and constitute interesting topics for future research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00819-x.

Acknowledgements We thank the Associate Editor and the two anonymous reviewers for their insightful comments which helped us to significantly improve this manuscript. This article is based on research using data obtained from www.projectdatasphere.org, which is maintained by *Project Data Sphere*. Neither *Project Data Sphere* nor the owner(s) of any information from the web site have contributed to, approved, or are in any way responsible for the contents of this article. Bakoyannis acknowledges funding support from Grants R21AI145662 and R01AI140854 from the National Institutes of Health. Bandyopadhyay acknowledges funding support from Grant P30CA016059 from the National Institutes of Health.

References

Aalen, O. O., Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics, 5*(3), 141–150.

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (2012). *Statistical models based on counting processes*. New York: Springer Science & Business Media.

Athreya, K. B., & Lahiri, S. N. (2006). *Measure theory and probability theory*. New York: Springer Science & Business Media.

Bakoyannis, G. (2020). Nonparametric tests for transition probabilities in nonhomogeneous Markov processes. *Journal of Nonparametric Statistics, 32*(1), 131–156.

Bakoyannis, G. (2021). Nonparametric analysis of nonhomogeneous multistate processes with clustered observations. *Biometrics, 77*(2), 533–546.

Begg, C. B., Larson, M. (1982). A study of the use of the probability-of-being-in-response function as a summary of tumor response data. *Biometrics, 38*(1), 59–66.

Bluhmki, T., Dobler, D., Beyersmann, J., Pauly, M. (2019). The wild bootstrap for multivariate Nelson–Aalen estimators. *Lifetime Data Analysis, 25*(1), 97–127.

Bluhmki, T., Schmoor, C., Dobler, D., Pauly, M., Finke, J., Schumacher, M., Beyersmann, J. (2018). A wild bootstrap approach for the Aalen–Johansen estimator. *Biometrics, 74*(3), 977–985.

Cai, T., Wei, L., Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika, 87*(4), 867–878.

Cameron, A. C., Gelbach, J. B., Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics, 90*(3), 414–427.

Campbell, M., Donner, A., Klar, N. (2007). Developments in cluster randomized trials and Statistics in Medicine. *Statistics in Medicine, 26*(1), 2–19.

Capasso, V., & Bakstein, D. (2015). *An introduction to continuous-time stochastic processes*. Basel: Birkhäuser.

Datta, S., 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 & Probability Letters, 55*(4), 403–411.

Datta, S., Satten, G. A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics, 58*(4), 792–802.

de Uña-Álvarez, J., Mandel, M. (2018). Nonparametric estimation of transition probabilities for a general progressive multi-state model under cross-sectional sampling. *Biometrics, 74*(4), 1203–1212.

de Uña-Álvarez, J., Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics, 71*(2), 364–375.

Ellis, S., Carroll, K. J., Pemberton, K. (2008). Analysis of duration of response in oncology trials. *Contemporary Clinical Trials, 29*(4), 456–465.

Fong, Y., Huang, Y., Lemos, M. P., Mcelrath, M. J. (2018). Rank-based two-sample tests for paired data with missing values. *Biostatistics, 19*(3), 281–294.

Glidden, D. V. (2002). Robust inference for event probabilities with non-Markov event data. *Biometrics, 58*(2), 361–368.

Kahan, B. C., Morris, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine, 31*(4), 328–340.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. New York: Springer Science & Business Media.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). Hoboken: John Wiley & Sons.

Liu, D., Kalbfleisch, J. D., Schaubel, D. E. (2011). A positive stable frailty model for clustered failure time data with covariate-dependent frailty. *Biometrics, 67*(1), 8–17.

Putter, H., 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*(7), 2081–2092.

Seaman, S., Pavlou, M., Copas, A. (2014). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine, 33*(30), 5371–5387.

Seaman, S. R., Pavlou, M., Copas, A. J. (2014). Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics, 70*(2), 449–456.

Shorack, G. R., & Wellner, J. A. (2009). *Empirical processes with applications to statistics*. Philadelphia: SIAM.

Studer, M., Struffolino, E., Fasang, A. E. (2018). Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure. *Sociological Methodology, 48*(1), 103–135.

Tattar, P. N., Vaman, H. (2014). The k-sample problem in a multi-state model and testing transition probability matrices. *Lifetime Data Analysis, 20*(3), 387–403.

Temkin, N. R. (1978). An analysis for transient states with application to tumor shrinkage. *Biometrics, 34*(4), 571–580.

Titman, A. C. (2015). Transition probability estimates for non-Markov multi-state models. *Biometrics, 71*(4), 1034–1041.

US Food and Drug Administration, et al. (2018). Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Federal Register*.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Vermorken, J. B., Stöhlmacher-Williams, J., Davidenko, I., Licitra, L., Winquist, E., Villanueva, C., Foa, P., Rottey, S., Skladowski, K., Tahara, M., et al. (2013). Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SPECTRUM): An open-label phase 3 randomised trial. *The Lancet Oncology, 14*(8), 697–710.

Zhang, H., Schaubel, D. E., Kalbfleisch, J. D. (2011). Proportional hazards regression for the analysis of clustered survival data from case-cohort studies. *Biometrics, 67*(1), 18–28.

Zhou, B., Fine, J., Latouche, A., Labopin, M. (2012). Competing risks regression for clustered data. *Biostatistics, 13*(3), 371–383.