

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <http://dx.doi.org/10.1007/s11205-018-1947-7>

Social Indicators Research

Ranking nursing homes' performances through a latent Markov model with fixed and random effects --Manuscript Draft--

Manuscript Number:	SOCI-D-17-01346R2
Full Title:	Ranking nursing homes' performances through a latent Markov model with fixed and random effects
Article Type:	S.I. : Socio-Economic indicators for performance evaluate & quality assess
Keywords:	Latent Markov models; Nursing homes; Performance assessment; Ranking construction
Corresponding Author:	Giorgio Eduardo Montanari Universita degli Studi di Perugia Dipartimento di Scienze politiche Perugia, ITALY
Corresponding Author's Institution:	Universita degli Studi di Perugia Dipartimento di Scienze politiche
Order of Authors:	Giorgio Eduardo Montanari Marco Doretti
Funding Information:	
Abstract:	<p>In this paper, we aim at ranking a set of nursing homes based on their ability in maintaining their residents' physical conditions as good as possible. In this respect, we propose a nursing home performance indicator, which is essentially a probability to avoid resident health status worsening. Specifically, latent Markov models with covariates and normally distributed continuous random effects are fitted to produce standardised six-month ahead transition matrices, upon which the aforementioned index is based. Nursing home effects on these transition matrices are modelled through fixed as well as random effects. The performance index is used to build two distinct rankings, one of which also accounts for the variability induced by the estimation process. In this framework, several rankings can be obtained by combining the model specification (fixed versus random effects), the kind of ranking and the number of latent states, which is the typical sensitivity parameter of latent Markov models. Our methodological approach is applied to a dataset which was gathered from a health protocol implemented in Umbria (Italy). Results for this data show a rather high degree of robustness, in the sense that the obtained rankings are almost the same.</p>

Noname manuscript No. (will be inserted by the editor)
--

Ranking nursing homes' performances through a latent Markov model with fixed and random effects

Received: date / Accepted: date

Abstract In this paper, we aim at ranking a set of nursing homes based on their ability in maintaining their residents' physical conditions as good as possible. In this respect, we propose a nursing home performance indicator, which is essentially a probability to avoid resident health status worsening. Specifically, latent Markov models with covariates and normally distributed continuous random effects are fitted to produce standardised six-month ahead transition matrices, upon which the aforementioned index is based. Nursing home effects on these transition matrices are modelled through fixed as well as random effects. The performance index is used to build two distinct rankings, one of which also accounts for the variability induced by the estimation process. In this framework, several rankings can be obtained by combining the model specification (fixed versus random effects), the kind of ranking and the number of latent states, which is the typical sensitivity parameter of latent Markov models. Our methodological approach is applied to a dataset which was gathered from a health protocol implemented in Umbria (Italy). Results for this data show a rather high degree of robustness, in the sense that the obtained rankings are almost the same.

Keywords Latent Markov models · Nursing homes · Performance assessment · Ranking construction

1 Introduction

In many countries, the public health system relies on the services provided by public and private structures, within a regulated framework. Clearly, proper tools to evaluate the performance of these structures are essential to policy makers to pursue efficiency, effectiveness and quality of health care services.

Address(es) of author(s) should be given

Nursing home (NH) care, which is one of the main services (Makai et al, 2014) together with the hospital system, is not an exception to this scheme. In this respect, in the United States it is a common practice to collect data for NH comparison and to rank the facilities according to the quality of resident care. Such comparisons are generally based on quality indicators derived from a single assessment of patients' health, like the presence or the absence of certain conditions. These indicators are then aggregated to the facility, regional and national level, and expressed as prevalence or incidence rates; see for example Arling et al (2005, 2007); Castle and Ferguson (2010); Phillips et al (1997) and references therein. Care quality evaluations are offered to interested people for a conscious choice and for stimulating the improvement of NH performances.

The above approach is also common in the evaluation of hospitals with respect to the outcome of patient treatments. However, in the case of NH care there are two aspects characterising the offered service. First, NH care is a long term facility, meaning that it is common for an elderly person entering an NH to spend there a considerable amount of time (rather often the rest of their life). Therefore, NH evaluation can be also carried out assuming a longitudinal perspective and developing some indicators of the ability to preserve resident health status as good as possible over time. Second, the outcome of interest, that is, resident health status, is an unobservable variable which is typically surveyed by a set of indicators. In this perspective, statistical NH evaluation methods dealing with both longitudinal data and the presence of unmeasured traits are of interest.

In this respect, latent Markov (LM) models (Wiggins, 1973; Bartolucci et al, 2013) are an appropriate tool. Indeed, LM models study an unobserved stochastic process which is assumed to evolve like a first-order Markov chain with a finite number of states. Though such a process is not directly observable, some proxies of it are assumed to be available. These proxies are typically referred to as the outcome variables, since they are manifestations of a common latent phenomenon. Specifically, in the application we consider, these outcome variables are the items of a suitable questionnaire.

The most relevant extensions of the basic LM model allow to include unit-level covariates either in the measurement model or in the latent model (Bartolucci et al, 2013; Vermunt et al, 1999). Applications of these models to the health care evaluation framework exist. For example, the former has been put forward to evaluate hospital efficiency (Pennoni and Vittadini, 2013). Conversely, the latter has been proposed to assess NH effects on patient health status, which can in principle be estimated by indicator variables representing group membership (Bartolucci et al, 2009; Montanari and Pandolfi, 2018). In the following, we refer to this second model as to the *fixed effect* LM model.

Alternative model specifications have been proposed which rely on random instead of fixed effects. Random effects have been introduced in various ways in the class of LM or hidden Markov models (Altman, 2007; Maruotti, 2011; Maruotti and Rocci, 2012). In this work, we limit our attention to the case where random effects aim at modelling the presence of some kind of clustering

in the data affecting the latent process, assuming a multilevel perspective. The literature about this specific model, which we call the *random effect* LM model, is quite limited: applications of it exist for medical (Koukounari et al, 2013) or educational datasets (Bartolucci et al, 2011) but not, to the best of our knowledge, in the context of health care evaluation.

In this work, we consider the NH care system of Umbria, a region of central Italy, and propose a methodology to evaluate the NH care quality by measuring the capability of each NH to keep their residents in good health conditions over time. Such a methodology is based on both fixed and random effect LM models and relies on data collected at the resident level. In particular, we extend results from previous works (Montanari et al, 2017a,b,c) and propose an NH performance index which can be interpreted as an overall probability to avoid the worsening of resident physical limitations. Such an index allows to compare and rank the NHs. Furthermore, we investigate the robustness of our rankings through a comparison between fixed and random effect LM models and two different ranking procedures. Such a robustness test is relevant in evaluation processes, as remarked also by other authors (see, for example, Gnaldi and Ranalli (2010, 2016) for a related discussion in the context of the evaluation of universities).

The paper is structured as follows. A brief description of the dataset we use is reported in Section 2, while the two competing models we consider, that is, the fixed effect and the random effect latent Markov model, are presented in Sections 3.1 and 3.2 respectively. Maximum likelihood estimation of the two models is discussed in Section 3.3, whereas in Section 3.4 we describe how the rankings we propose are built. Model results and NH rankings for the data at hand are shown in Section 4, while in Section 5 some final remarks are given.

2 The LTCF dataset

Our analysis is based on data collected under the Long Term Care Facilities Protocol (LTCF), which has been implemented by the regional government of Umbria since 2010. Data are gathered by administering NH residents, approximately every six months, a questionnaire belonging to an internationally validated tool termed *Suite interRAI* instruments (Hirdes et al, 2008; Kim et al, 2015).

The *interRAI* questionnaire in use is formed by several items referring to different domains of the health status (cognitive conditions, physical limitations, auditory and view fields, incontinence, *etc.*). In this work, we only focus on the section concerning the physical limitations, which are measured by 10 items referring to the so-called *Activities of Daily Living* (ADL) as in Montanari et al (2017a). Items in this section quantify residents' difficulties in every-day activities like washing, getting dressed, walking, eating, using the WC, maintaining personal hygiene, *etc.* The higher the response level, the higher the difficulty in the activity and the dependence on the assistance provided by other people because of the lack of autonomy. Specifically, an ordinal

scale with 6 levels (labelled from 1=“independent” to 6=“totally dependent”) is adopted.

Focusing on a single section of the questionnaire is a limitation of our approach, since other relevant domains of the health status are ignored. In our specific case - where attention is confined to the ADL section - one has thus to bear in mind that the latent trait being modelled is narrowed to residents’ *physical* limitations. However, this approach allows to deal with a one-dimensional latent variable, which can be reasonably assumed to have an ordinal nature. This setting allows us to rank the NHs with respect to their ability in preserving resident physical autonomy. Clearly, similar rankings can be built for other health domains.

In the LTCF dataset, some individual covariates are also available. Specifically, age and gender are measured at every time occasion, while the temporal distance (in days) from the previous measurement is present, clearly, from the second occasion onwards. Age is obviously correlated to the worsening of physical limitations, and so is gender (Montanari et al, 2017a). Temporal distance between observations is relevant since the data show some variability in the time intervals between occasions, which is likely to depend on resident health status. Typically, observations are anticipated with respect to the canonical six-month distance in the presence of a significant change in residents’ conditions. Conversely, delayed observations occur in the case of stable conditions.

The longitudinal dataset we consider refers to the years 2012 and 2013, and contains 3924 observations for 1292 residents distributed in 47 NHs. Table 1 contains means and standard deviations for the labels of the ADL item categories and for the individual covariates, across the whole set of observations, as well as some summary statistics for the distributions of the 47 NHs according to their number of patients and observations.

As already mentioned, the *interRAI* questionnaire is administered roughly every six months. Since our dataset covers two years, ideally each of the 1292 residents should have four measurement occasions. However, the dataset at hand comprises only 3924 rather than 5168 observations due to intermittent missingness (that is, missing observations before the last available) and dropout. The former is relatively rare and has unknown causes. Therefore, it is assumed to occur at random. Conversely, dropout has several causes, the most relevant of which is resident death. Clearly, a probabilistic dependence exists between dropout due to death and the outcome variables (*i.e.*, the ADL indicators). Indeed, the former is likely to be associated to a worsening in the values of the latter. Therefore, such a missing data mechanism is non-ignorable, and has to be somehow modelled in order to avoid biased estimates. An account on how dropout due to death is handled is given in Section 3. On the other hand, dropout due to other reasons - such as discharge, transfer to other structures or similar - is far less frequent and, like intermittent missingness, is assumed to occur at random, since a clear connection with the ADL values cannot be established.

ADL items & covariates						
	description	mean	std. dev.			
Y_1	Use of the shower stall/bath tub	4.78	1.44			
Y_2	Personal hygiene	4.53	1.63			
Y_3	Dressing the upper part of the body	4.38	1.77			
Y_4	Dressing the lower part of the body	4.62	1.68			
Y_5	Walking	4.08	2.06			
Y_6	Locomotion	3.99	2.09			
Y_7	Transfer to the WC	4.19	1.99			
Y_8	WC use	4.38	1.91			
Y_9	Bed mobility	3.56	2.04			
Y_{10}	Eating	2.76	2.02			
X_1	age (years)	82.30	10.37			
X_2	gender (1=female)	0.72	0.45			
X_3	distance from previous obs. (days)	187.05	48.65			
nursing homes						
	min	1Q	median	mean	3Q	max
# of patients	1.00	9.50	19.00	27.49	39.00	96.00
# of obs.	1.00	27.50	62.00	83.49	114.50	309.00

Table 1 Summary statistics for the LTCF data: means and standard deviations for the ADL items and the individual covariates; range, mean and quartiles of the number of patients/observations across NHs.

3 Latent Markov models and ranking construction

3.1 Fixed effect model

In this section, we present the fixed effect LM model developed for the data at hand to evaluate NH performances. Assume we have n independent sample units, $i = 1, \dots, n$, that in our case are the 1292 NH residents. Let $\mathbf{Y}_i^{(t)}$, $\mathbf{X}_i^{(t)}$, and $\mathbf{Z}_i^{(t)}$ respectively denote the item response vector, the individual covariate vector and the NH membership indicator vector of unit i at occasion $t = 1, \dots, T_i$. Specifically, the vector $\mathbf{Y}_i^{(t)} = (Y_{i1}^{(t)}, \dots, Y_{iJ}^{(t)})$ collects $J = 10$ univariate categorical items. As stated in Section 2, here each item has a constant number $c = 6$ of response categories, labelled from 1 to 6, though in principle such a number can be different for every item. Further, the set of individual covariates in $\mathbf{X}_i^{(t)}$ can vary across time; see Section 2. Finally, $\mathbf{Z}_i^{(t)}$ contains $H = 47$ indicator variables. The one corresponding to the NH unit i belongs to is set equal to one, while all the others are set to zero. Despite this general notation, in our application $\mathbf{Z}_i^{(t)}$ is indeed time-invariant, since for our purposes the few residents that leave the NH hosting them in the first occasion are treated as dropouts. Each sample unit has its own number of measurement occasions $T_i \leq T = 4$, and individual vectors can be collected across time in the vectors $\mathbf{Y}_i = (\mathbf{Y}_i^{(1)}, \dots, \mathbf{Y}_i^{(T_i)})$, $\mathbf{X}_i = (\mathbf{X}_i^{(1)}, \dots, \mathbf{X}_i^{(T_i)})$ and $\mathbf{Z}_i = (\mathbf{Z}_i^{(1)}, \dots, \mathbf{Z}_i^{(T_i)})$.

We assume that, at occasion t , $\mathbf{Y}_i^{(t)}$ depends on an unobserved discrete latent variable $V_i^{(t)}$ - which in our application reflects resident i 's level of physical limitations - with k levels. We recall that in this context these levels are ordered, from no limitation at all (level 1) to the most severe limitations (level k). A first-order Markov chain is assumed to govern the whole latent process $\mathbf{V}_i = (V_i^{(1)}, \dots, V_i^{(T_i)})$.

Like in every LM model, the parameters of interest can be divided into three groups: i) conditional response probabilities, ii) initial probabilities, and iii) transition probabilities. The first are the probabilities of observing specific item categories from a unit in a given latent state. The second are the probabilities of being in a latent state at the first measurement occasion, whereas the third are the probabilities of moving towards a latent state given one was in a specific latent state at the previous measurement occasion. Although covariates may affect each group of parameters (Bartolucci et al, 2017), here they are assumed to affect parameters in the second and third group only. In fact, in our application, while age and gender are clearly predictors of health status, their effect on the measurement model is questionable. Furthermore, our purpose is to identify groups of elders whose physical limitations are stable or improve over time. It follows that conditional response probabilities are assumed to be constant across time and to depend on unit i only through its latent state v . Therefore, we set

$$P(Y_{ij}^{(t)} = y | V_i^{(t)} = v) = \phi_{jyv}$$

($j = 1, \dots, J$, $y = 1, \dots, c$, $v = 1, \dots, k$, $i = 1, \dots, n$, $t = 1, \dots, T_i$). Furthermore, we assume that, conditionally on $V_i^{(t)}$, each item $Y_{ij}^{(t)}$ is independent of any other variable in the model.

Conditional response probabilities can be parametrised in many different ways in order to reduce the number of free parameters that have to be estimated. Because of the ordinal nature of the response items and of the latent trait, a global logit parametrisation (see Bartolucci et al., 2013; p. 30)

$$\log \frac{\phi_{jm+1v} + \dots + \phi_{jcv}}{\phi_{j1v} + \dots + \phi_{jmv}} = \tau_{jm} + \delta_v \quad (1)$$

($j = 1, \dots, J$; $m = 1, \dots, c - 1$; $v = 1, \dots, k$) can be set. In detail, the parameters τ_{jm} form J sequences of thresholds such that $\tau_{j1} > \dots > \tau_{jc-1}$ for $j = 1, \dots, J$, whereas the parameters δ_v are increasing with the latent state, that is, $\delta_1 < \dots < \delta_k$, with δ_1 set to 0 for model identifiability.

The latent model is specified by the conditional initial probabilities

$$\pi_i^{(1)}(v) = P(V_i^{(1)} = v | \mathbf{X}_i^{(1)} = \mathbf{x}_i^{(1)}, \mathbf{Z}_i^{(1)} = \mathbf{z}_i^{(1)})$$

($v = 1, \dots, k$; $i = 1, \dots, n$) and by the conditional transition probabilities

$$\pi_i^{(t)}(v|\bar{v}) = P(V_i^{(t)} = v | V_i^{(t-1)} = \bar{v}, \mathbf{X}_i^{(t)} = \mathbf{x}_i^{(t)}, \mathbf{Z}_i^{(t)} = \mathbf{z}_i^{(t)})$$

($\bar{v}, v = 1, \dots, k$; $i = 1, \dots, n$; $t = 2, \dots, T_i$). Contrary to conditional response probabilities, these quantities are indexed by i and t as they depend

on individual-specific covariates. However, the dependence on $\mathbf{x}_i^{(t)}$ and $\mathbf{z}_i^{(t)}$ ($t = 1, \dots, T_i$) is suppressed in $\pi_i^{(1)}(v)$ and $\pi_i^{(t)}(v|\bar{v})$ to ease notation. Again, since latent states are ordered, a global logit link for the regression equations of these probabilities is assumed, that is

$$\log \frac{\pi_i^{(1)}(v+1) + \dots + \pi_i^{(1)}(k)}{\pi_i^{(1)}(1) + \dots + \pi_i^{(1)}(v)} = \xi_v + \mathbf{x}_i^{(1)}\boldsymbol{\beta}_0 + \mathbf{z}_i^{(1)}\boldsymbol{\beta}_1 \quad (2)$$

($v = 1, \dots, k-1$; $i = 1, \dots, n$), and

$$\log \frac{\pi_i^{(t)}(v+1|\bar{v}) + \dots + \pi_i^{(t)}(k|\bar{v})}{\pi_i^{(t)}(1|\bar{v}) + \dots + \pi_i^{(t)}(v|\bar{v})} = \psi_{\bar{v}} + \omega_v + \mathbf{x}_i^{(t)}\boldsymbol{\gamma}_0 + \mathbf{z}_i^{(t)}\boldsymbol{\gamma}_1 \quad (3)$$

($\bar{v} = 1, \dots, k$; $v = 1, \dots, k-1$; $i = 1, \dots, n$; $t = 2, \dots, T_i$). Like in (1), in (2) and (3) there are sequences of ordered thresholds. Specifically, we have $\xi_1 > \dots > \xi_{k-1}$ and $\omega_1 > \dots > \omega_{k-1}$. On the contrary, the sequence $\psi_{\bar{v}}$ ($\bar{v} = 1, \dots, k$) does not need to be ordered, though the constraint $\psi_1 = 0$ must be imposed for model identifiability. The column vectors $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ measure, respectively, the individual covariate effects and the NH effects on the initial probabilities, whereas $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$ measure the analogous effects on the transition probabilities. All these effect vectors do not change across the logit equations. To ensure model identification, a corner point parametrisation is adopted for $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$: the first NH is taken as reference and its coefficients are set to zero, so that all other coefficients have to be interpreted as effect differences from this reference NH (given the values of individual covariates) on the linear predictor scale.

The NH effects in $\boldsymbol{\beta}_1$ account for the heterogeneity unexplained by age and gender in the initial probabilities of the latent states. As such, they can be reasonably imputed to the different admission policies of residents in the NHs. Similarly, the NH effects in $\boldsymbol{\gamma}_1$ account for the heterogeneity not explained by age, gender and temporal distance between observations in the transition probability matrices of residents. A positive/negative value means that, conditional on the latent state at time $t-1$ and other covariates at time t , the probability of a transition toward a higher latent state (*i.e.*, more serious physical limitations) at time t is higher/lower than the same probability for the reference NH. In this sense, these effects can be taken as indicators of the quality of the care provided by the NHs.

To tackle the presence of non-ignorable dropout as described in Section 2, the model is extended as follows. For each outcome variable, we define an additional response category $c+1$, and we augment the data trajectories for residents who die after the t -th occasion, letting $Y_{hij}^{(s)} = c+1$ for $s = t+1, \dots, T$. In our application, the augmented dataset we obtain after this procedure has 4746 observations. Furthermore, we introduce an additional absorbing latent state $k+1$ corresponding to death. As a consequence, a number of additional initial, transition and conditional response probabilities are introduced, and

some of these have to be set to specific values. In detail, for $j = 1, \dots, J$, $v = 1, \dots, k$ and $t = 2, \dots, T$, we set:

- i) $\pi_i^{(1)}(k+1) = 0$ ($i = 1, \dots, n$): no one can be in the latent state associated to death in the first occasion;
- ii) $\pi_i^{(t)}(k+1|k+1) = 1$ ($i = 1, \dots, n$): no one can revert to other states from death (that is, death is an *absorbing* state);
- iii) $\phi_{j(c+1)v} = 0$: the additional response category cannot be observed if one is not dead.

Notice that constraints i) and ii) have a conceptual nature, whereas iii) is linked to the data expansion we performed. Overall, $\pi_i^{(t)}(k+1|1), \dots, \pi_i^{(t)}(k+1|k)$ are the only new probabilities that have to be estimated to account for dropout due to death. These are the probabilities that a resident in a generic latent state v ($v = 1, \dots, k$) at occasion $t-1$ will be dead at occasion t . It is worth to remark that this additional model feature implies the estimation of a single additional free parameter - that is, the threshold ω_k - that has to satisfy the previously mentioned inequality constraint for transition probabilities, *i.e.*, $\omega_1 > \dots > \omega_{k-1} > \omega_k$.

3.2 Random effect model

In the model of the above section, β_1 and γ_1 represent the NH fixed effects on the initial and transition probabilities. In this section, we introduce the random effect LM model, which, as stated in Section 1, can be essentially seen as a multilevel model. Although the overall model structure is similar to that of the fixed effect LM model, there are some notable distinctions. The major difference is that group membership is no more accounted for by introducing indicator covariates reflecting NH membership (*i.e.*, the indicator variables in \mathbf{Z}_i), but it is an intrinsic feature of the model structure. Specifically, since each of the n sample units belongs to a cluster (an NH), the double index hi is used to identify units, with h denoting clusters ($h = 1, \dots, H$) and i denoting units within clusters ($i = 1, \dots, n_h$), where n_h is the sample size of the h -th cluster and $\sum_{h=1}^H n_h = n$. Apart from this difference, the notation for the outcome variables, the covariates and the latent process is the same as in Section 3.1. Clearly, one has to bear in mind that another level of aggregation for random vectors is possible: for example, $\mathbf{X}_h = (\mathbf{X}_{h1}, \dots, \mathbf{X}_{hn_h})$ denotes the collection of all individual covariate vectors of units in cluster h (with \mathbf{Y}_h and \mathbf{V}_h meaning the same for the outcome variables and the latent process, respectively).

As typical in multilevel models, different clusters are assumed marginally independent, whereas units in the same cluster are not. Here, we assume that within-cluster independence holds conditionally on a cluster-specific vector of random effects $\mathbf{U}_h = (R_h, S_h)$, where R_h affects the initial probabilities and S_h affects the transition probabilities. Therefore, letting

$$\pi_{hi}^{(1)}(v) = P(V_{hi}^{(1)} = v | \mathbf{X}_{hi}^{(1)} = \mathbf{x}_{hi}^{(1)}, R_h = r_h)$$

$(v = 1, \dots, k; h = 1, \dots, H; i = 1, \dots, n_h)$ and

$$\pi_{hi}^{(t)}(v|\bar{v}) = P(V_{hi}^{(t)} = v | V_{hi}^{(t-1)} = \bar{v}, \mathbf{X}_{hi}^{(t)} = \mathbf{x}_{hi}^{(t)}, S_h = s_h)$$

$(v, \bar{v} = 1, \dots, k; h = 1, \dots, H; i = 1, \dots, n_h; t = 2, \dots, T_{hi})$, the equivalent of Equations (2) and (3) for the random effect LM model is given by

$$\log \frac{\pi_{hi}^{(1)}(v+1) + \dots + \pi_{hi}^{(1)}(k)}{\pi_{hi}^{(1)}(1) + \dots + \pi_{hi}^{(1)}(v)} = \tilde{\xi}_v + \mathbf{x}_{hi}^{(1)} \tilde{\boldsymbol{\beta}}_0 + \sigma_r r_h \quad (4)$$

$(v = 1, \dots, k-1; h = 1, \dots, H; i = 1, \dots, n_h)$ and

$$\log \frac{\pi_{hi}^{(t)}(v+1|\bar{v}) + \dots + \pi_{hi}^{(t)}(k|\bar{v})}{\pi_{hi}^{(t)}(1|\bar{v}) + \dots + \pi_{hi}^{(t)}(v|\bar{v})} = \tilde{\psi}_{\bar{v}} + \tilde{\omega}_v + \mathbf{x}_{hi}^{(t)} \tilde{\boldsymbol{\gamma}}_0 + \sigma_s s_h \quad (5)$$

$(v = 1, \dots, k-1; \bar{v} = 1, \dots, k; h = 1, \dots, H; i = 1, \dots, n_h; t = 2, \dots, T_{hi})$, where the symbol \sim is added above the parameters $(\xi_v, \psi_{\bar{v}}, \omega_v, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ to distinguish them from the respective parameters in the fixed effect model, although their interpretation remains unchanged. The NH effects in $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ of the fixed effect model are replaced by the random effects $\sigma_r r_h$ and $\sigma_s s_h$, respectively.

To complete the model specification, a distribution needs to be assigned to the random vector \mathbf{U}_h . Specifically, we assume it to follow a standard bivariate normal distribution and denote by ρ the correlation coefficient of this distribution. Thus, the coefficients σ_r and σ_s in (4) and (5) can be interpreted as the standard deviations of the overall NH effects $\sigma_r R_h$ and $\sigma_s S_h$. Normality is by far the most popular choice for random effects within this class of models, the main reason lying in the fact that normal integrals can be efficiently approximated by quadrature methods; see Section 3.3. Although in principle it might be questioned, such an assumption does not seem unrealistic in the application at hand. Indeed, NH effects on both initial and transition probabilities can be thought of as average effects summarising a number of distinct components, so that normality is justified overall.

Under this formulation, the expected posterior random effects $\sigma_s E(S_h | \mathbf{Y}_h)$ allow to rank the NHs according to their probabilities of transition toward higher degrees of physical limitations. Indeed, these probabilities are monotone functions of the random effects. This is how the longitudinal structure of the available data is exploited, albeit with some provisos discussed in Section 4, for evaluation purposes. Similar approaches, though in a cross-sectional setting, are quite popular within the evaluation framework in many fields, ranging from hospitals (Vittadini and Minotti, 2005) to educational institutions (Grilli and Rampichini, 2009; Rampichini et al, 2004).

3.3 Maximum likelihood estimation

Maximum likelihood estimation is performed for both models. However, two different procedures are adopted which are briefly described in this section. In

what follows, the vectors of all model parameters for the fixed and the random effect models are denoted respectively by $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_r$, whereas their maximum likelihood estimates by $\hat{\boldsymbol{\theta}}_f$ and $\hat{\boldsymbol{\theta}}_r$. Specifically, letting p be the total number of elements of $\tilde{\boldsymbol{\beta}}_0$ and $\tilde{\boldsymbol{\gamma}}_0$, $\boldsymbol{\theta}_r$ contains $4k + p + J(c - 1)$ parameters, whereas $\boldsymbol{\theta}_f$ has $2(H - 1) - 3$ additional parameters since both $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}_1$ have $H - 1$ components, while the random effect model accounts for NH effects by means of three parameters only.

For the fixed effect LM model, the Expectation-Maximisation (EM) algorithm (Dempster et al, 1977) is implemented to maximise the model log-likelihood

$$\ell(\boldsymbol{\theta}_f) = \sum_{i=1}^n \log P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i). \quad (6)$$

The EM algorithm is the most popular estimation tool for LM models in their classical formulation; see Bartolucci et al (2013) for a detailed overview. Such a method consists in iteratively alternating two steps. Starting from a random guess for $\boldsymbol{\theta}_f$, in the first step (E-step) the conditional expectation of the *complete* data log-likelihood

$$\ell^*(\boldsymbol{\theta}_f) = \sum_{i=1}^n \log P(\mathbf{Y}_i = \mathbf{y}_i, \mathbf{V}_i = \mathbf{v}_i | \mathbf{X}_i = \mathbf{x}_i, \mathbf{Z}_i = \mathbf{z}_i) \quad (7)$$

given the data is computed. Then, this conditional expected complete data log-likelihood is maximised to update the parameter vector estimate of $\boldsymbol{\theta}_f$ (M-step). This scheme is repeated until convergence to obtain the final estimate $\hat{\boldsymbol{\theta}}_f$. Given our model parametrisation, the M-step involves a constrained maximisation procedure in order to obtain the ordered threshold sequences as described in Section 3.1. Specific R code was developed to implement the whole algorithm, with a Fortran interface to speed computation as much as possible.

For the continuous random effect LM model, EM estimation is much more cumbersome. Therefore the log-likelihood

$$\ell(\boldsymbol{\theta}_r) = \sum_{h=1}^H \log P(\mathbf{Y}_h = \mathbf{y}_h | \mathbf{X}_h = \mathbf{x}_h) \quad (8)$$

is maximised directly. Given the model assumptions, each cluster-specific component $P(\mathbf{Y}_h = \mathbf{y}_h | \mathbf{X}_h = \mathbf{x}_h)$ in (8) can be written as

$$\int_{\mathbb{R}^2} \left[\prod_{i=1}^{n_h} P(\mathbf{Y}_{hi} = \mathbf{y}_{hi} | \mathbf{X}_{hi} = \mathbf{x}_{hi}, \mathbf{U}_h = \mathbf{u}_h) \right] \phi_\rho(\mathbf{u}_h) d\mathbf{u}_h, \quad (9)$$

where $\phi_\rho(\mathbf{u}_h)$ denotes the density function at \mathbf{u}_h of a standard bivariate normal distribution with correlation ρ , which is the correlation between the two components of \mathbf{U}_h . In practice, the integral in (9) is approximated by a Gauss-Hermite quadrature method, and direct maximisation is performed by using

the BFGS algorithm (Fletcher, 1987), an iterative optimisation algorithm readily available in R. We highlight that the explicit computation of the gradient of (8) is needed to run the BFGS algorithm. Again, a combination of R and Fortran routines is used to calculate such a gradient vector as well as the log-likelihood in (8) itself. Suitable parameter transformations are introduced to ensure the estimated standard deviations $\hat{\sigma}_r$ and $\hat{\sigma}_s$ be positive and $\hat{\rho}$ lie between -1 and 1. Conversely, threshold orderings are met in the final estimate $\hat{\theta}_r$ without imposing any constraint, since a sensible starting value for the BFGS algorithm is chosen to ease convergence (see Montanari et al (2017a) for the details).

For both models, Hessian matrices are needed to obtain variance-covariance matrices (and standard errors) of $\hat{\theta}_f$ and $\hat{\theta}_r$. For the random effect model, the Hessian matrix is directly returned by the BFGS algorithm, whereas for the fixed effect model the equivalence between the derivative of (6) and that of the expected complete log-likelihood is exploited (Oakes, 1999). In detail, the latter is first computed by implementing an E-step from the solution $\hat{\theta}_f$ and then further derived to obtain the Hessian matrix.

Finally, it is worth to discuss computational times in relation to the application considered in this work. To this end, we recall that the dataset at hand consists of 4746 observations referring to 1292 units, which are divided in 47 clusters. Further, there are 10 outcome variables with 6 categories each and 3 covariates; see Section 2. For this data burden, computational times are reasonable for the EM algorithm, while the BFGS algorithm takes, on a standard machine, more than one day for each model. Clearly, in any case the computational complexity increases with the number of latent states k .

3.4 The NH performance indicator and the rankings

Service performance evaluation is a standing field of research, which may have many different objects. Here, interest lies in NH care quality evaluation, where - as mentioned in Section 1 - most methods are typically based on indicators computed using cross-sectional data (Castle and Ferguson, 2010). A similar strategy has been pervading also other areas, where multilevel cross-sectional models have become a widely employed tool (Vittadini and Minotti, 2005; Rampichini et al, 2004). Attempts to evaluate services on the basis of the longitudinal evolution of the outcomes have also been made; see for example Pennoni and Vittadini (2013) and Colombi et al (2017) in the field of hospital efficiency. However, these approaches rely on data at the hospital level. Conversely, we aim at evaluating health services from a patient-based standpoint, so methods tailored to longitudinal resident-level data sources are needed.

Motivated by these considerations, we now show how output from the fixed effect and random effect (multilevel) LM models is employed to assess NH contributions to changes in resident physical health status. Specifically, our approach is based upon the idea that smaller probabilities of transition toward

a higher level of physical limitations are associated to a better overall quality of care in terms of assistance, prevention of traumatic events and provided treatment of chronic diseases. Therefore, a plausible performance indicator can be based on the transition matrices estimated via the aforementioned models. Because 180 days is the canonical distance that approximately separates two consecutive measurements for each resident (see Section 3.1), the transitions matrices we deal with refer to this time interval. Further, it is reasonable to assume that 180 days represent a time interval wide enough to allow NH practices to be evaluated.

As age and gender also affect the latent state transitions, 180-day ahead transition matrices are individual-specific. Therefore, an aggregation method has to be implemented in order to come up with a single overall transition matrix for each NH. Taking the simple average of the matrices of residents in the same NH would not be appropriate. Indeed, NHs have different compositions with respect to age and gender, so NH comparisons based on such average matrices would be also affected by these differences. To overcome this issue, we rely on the standard population method (Kitagawa, 1964), which is widely used in demography. In our context, such a method implies each NH-specific overall transition matrix be obtained by averaging across the same set of units (specifically, the whole set of residents). In detail, for each NH a fictitious dataset is constructed as if all residents belonged to that NH. In the fixed effect model, this is done by suitably modifying the \mathbf{Z}_i vectors, whereas in the random effect model the same can be achieved by assigning every resident the estimated posterior expected effect $\hat{\sigma}_s E(S_h | \mathbf{Y}_h)$ for NH h .

The procedure described above produces NH-specific standard transition matrices depending on the estimated NH effect only. Clearly, these matrices can be used to obtain a performance-based ranking of the NHs. While fixed effect models have already been proposed to a similar end (Bartolucci et al, 2009; Montanari and Pandolfi, 2018), the adoption of the multilevel LM model for ranking purposes is rather innovative. This is due to the fact that previously fitted multilevel LM models include random effects with a discrete distribution (Bartolucci et al, 2011; Koukounari et al, 2013), which can provide at most a clustering rather than a ranking of second level units. In this sense, the introduction of continuous random effects plays a key role in the whole evaluating framework.

For a k -state model, the estimated 180-day ahead standard transition matrix for the h -th NH takes the form

$$\hat{\mathbf{\Pi}}_h = \begin{pmatrix} \hat{\pi}_h(1|1) & \dots & \hat{\pi}_h(k|1) & \hat{\pi}_h(k+1|1) \\ \vdots & \ddots & \vdots & \vdots \\ \hat{\pi}_h(1|k) & \dots & \hat{\pi}_h(k|k) & \hat{\pi}_h(k+1|k) \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where all the rows sum to one and the last row refers to the additional latent state introduced to handle dropout due to death, as detailed in Section 3. Since the latent states $1, \dots, k$ are ordered - from the least to the most serious

physical limitations - a summary measure of the NH capability to avoid the worsening of resident physical limitations can be given by

$$\hat{p}_h = \frac{1}{k} \sum_{\bar{v}=1}^k \sum_{v=1}^{\bar{v}} \hat{\pi}_h(v|\bar{v}). \quad (10)$$

This quantity corresponds to the average of the cumulative row probabilities, up to the main diagonal of $\hat{\mathbf{\Pi}}_h$, excluding the last row, and it can be interpreted as the probability that a randomly selected resident, once assigned to NH h , does not worsen the physical limitations within 180 days. Clearly, higher values of \hat{p}_h are associated to better NH performances in terms of quality of the care provided.

Using the performance index \hat{p}_h , we introduce two ranking methods. The first method simply ranks the NHs according to the value of \hat{p}_h , from the highest to the lowest. However, such a procedure ignores the uncertainty deriving from the estimation process.

To take into account the uncertainty in the estimates, hence the uncertainty in the NH rank, we propose a second procedure where the variability of \hat{p}_h is also relevant. From a conceptual standpoint, this approach is similar to others relying on Monte Carlo methods (Marozzi and Bolzan, 2016a,b). The construction of the second ranking involves three steps:

- i) the computation of the estimated variances of \hat{p}_h , for $h = 1, \dots, H$;
- ii) the computation of the 95% overlap intervals (OIs), under the normality assumption for \hat{p}_h , for multiple comparison (Goldstein and Healy, 1995; Afshartous and Preston, 2010);
- iii) the construction of a synthetic index based on the OIs to obtain the ranking.

The first step entails to compute the approximate standard errors of the elements of $\hat{\mathbf{\Pi}}_h$, which can be obtained via the delta method starting from the standard errors of $\hat{\boldsymbol{\theta}}_f$ and $\hat{\boldsymbol{\theta}}_r$. Computing these standard errors is rather complex. Here, we make the simplifying assumption that the elements of $\hat{\mathbf{\Pi}}_h$ are uncorrelated. Under this assumption, the estimated variance of \hat{p}_h is computed as

$$\hat{V}(\hat{p}_h) = \frac{1}{k^2} \sum_{\bar{v}=1}^k \sum_{v=1}^{\bar{v}} \hat{V}(\hat{\pi}_h(v|\bar{v})), \quad (11)$$

with $\hat{V}(\hat{\pi}_h(v|\bar{v}))$ denoting the estimated variance of $\hat{\pi}_h(v|\bar{v})$.

Once the variances of the \hat{p}_h estimates are calculated, pairwise comparisons can be performed to assess whether statistically significant differences exist among the values of \hat{p}_h . In the second step, in order to operate multiple comparisons at the right level of significance, we follow the approach of Goldstein and Healy (1995) and compute 95% OIs setting the critical level of each 95% OI equal to the average of the critical levels of all pairwise comparisons.

Then, for any given couple of NHs, say h and h' , we say that h is significantly better than h' when the lower OI limit of h is greater than the upper

OI limit of h' ; conversely, when the upper OI limit of h is smaller than the lower OI limit of h' , we say that h is significantly worse than h' .

Finally, the synthetic index for the h -th NH is computed as

$$i_h = \frac{1}{2} \left(\frac{W_h - B_h}{H - 1} + 1 \right),$$

where W_h is the number of NHs performing significantly worse than NH h , and B_h is the number of NHs performing significantly better than NH h . Like \hat{p}_h , the i_h index ranges between 0 and 1, with higher values associated to better performances. Note that two or more NHs might have the same value of i_h : as typical, in the ranking such ties are handled by assigning the average position among those into play. It is also worth to remark that in this second ranking procedure the position of an NH depends on the value of both \hat{p}_h and $\hat{V}(\hat{p}_h)$. If the latter is high, then the rank is uncertain and W_h and B_h might be very small, even zero. In such a case the NH would be classified approximately in the middle of the ranking.

4 Results for the LTCF dataset

In this section, we give details about the estimated LM models for the LTCF data. Specifically, we first highlight how model selection is performed. Then, we briefly discuss the main results for the final models and implement the ranking procedures proposed in the previous section. Prior to this discussion, it is important to clarify that the comparison between the fixed and random effects models cannot involve all the 47 NHs observed in the LTCF dataset. Specifically, for the fixed effect model, estimates for only 41 NHs are available. This is due to a drawback of the EM algorithm, in which estimation of some NH effects might be highly unstable, if not completely unfeasible, when just a few units carry information about them. To overcome this issue, we have removed the six NHs with less than ten observations. Conversely, for the random effect model, the BFGS algorithm does not suffer from this problem, so results for all the 47 NHs are available. However, in order to compare the two models, we focus only on the 41 largest NHs. Henceforth, we set $H = 41$ and use the subscript $h = 1, \dots, 41$ to index the NHs in this restricted set.

For both models, model selection is essentially concerned with the choice of the number of latent states k . To this end, a sensitivity analysis is performed where a number of different values for k are specified. In detail, for the fixed effect model we let k range between 2 and 10, whereas for the random effect model we set $k \in \{2, \dots, 7\}$, since higher values result in prohibitive computational times (see Section 3.3). Formal indices based on the penalised log-likelihood like the Bayesian Information Criterion (Schwarz, 1978) are initially adopted as a selection strategy. However, since it is known that these indices often tend to select models with too many latent states (Bacci et al, 2014), we take also alternative model selection criteria into account (Pohle

k	RE model			FE model		
	ℓ	$\#par$	BIC	ℓ	$\#par$	BIC
2	-46,017.8	63	92,487.0	-46,206.8	142	92,875.4
3	-40,704.1	67	81,888.2	-40,911.8	146	82,277.9
4	-38,460.7	71	77,430.1	-38,675.6	150	77,818.0
5	-37,736.6	75	76,010.5	-37,961.4	154	76,401.8
6	-37,413.8	79	75,393.6	-37,646.2	158	75,784.0
7	-37,165.0	83	74,924.5	-37,407.0	162	75,318.0

Table 2 Log-likelihood, number of parameters and BIC values for random (RE) and fixed (FE) effects LM fitted models.

et al, 2017). Specifically, we consider the overall interpretability of the latent states, and how sharply each model classifies a posteriori sample units in these latent states. Overall, for both the random and the fixed effect model, the one with $k = 5$ latent states is selected as the final one; see Montanari et al (2017a,b,c) for further details. Here, in Table 2 we report the log-likelihood, the number of parameters and the BIC index for the fixed effects and the random effects multilevel LM fitted models.

The analysis of normal ordinary pseudo-residuals (Zucchini and MacDonald, 2009, Ch. 6) does not raise concerns related to model fitting for any item. However, to prevent model misspecification we also investigate the presence of a quadratic effect of age in the latent model. Such an effect does not prove to be significant and is therefore removed.

The latent states identified by the two models are essentially the same. In Table 3, we report the latent state profiles in terms of estimated conditional mean values of the ADL items listed in Section 2. These refer to the fixed effect model, with those of the random effect model being almost identical. We recall that, due to the ordinal nature of the latent trait, these states are ordered, with the first one including residents in the best physical conditions. In summary, residents in the first latent state show some difficulties in a specific set of activities which includes taking a shower, maintaining the personal hygiene and getting dressed. Members of the second state experience the same difficulties, though to a greater extent, together with some initial problems in walking and using the WC. Residents in latent state 3 require rather intensive assistance for all the actions but eating and moving in and out the bed, which is instead more problematic for residents in the fourth state. Finally, the last state contains residents who are almost totally unable to carry out the surveyed ADLs.

An important point that has to be discussed before showing the values of the index \hat{p}_h and the rankings based on it concerns the correlation between the NH effects on initial and transition probabilities. Indeed, the former reflects NHs' tendency to admit residents in better or worse conditions. Therefore, to prevent any form of adverse selection in the resident admission process, a proper evaluation scheme is required to produce results not affected by the case-mix, that is, the different complexity each NH has to deal with at the

item	description	latent state v				
		1	2	3	4	5
Y_1	Use of the shower stall/bath tub	2.18	3.38	4.34	5.38	5.96
Y_2	Personal hygiene	1.62	3.10	4.11	5.17	5.94
Y_3	Dressing the upper part of the body	1.32	2.70	3.85	5.05	5.93
Y_4	Dressing the lower part of the body	1.50	3.10	4.24	5.35	5.96
Y_5	Walking	1.04	1.65	3.22	4.94	5.93
Y_6	Locomotion	1.03	1.53	3.05	4.83	5.92
Y_7	Transfer to the WC	1.06	1.95	3.54	5.03	5.93
Y_8	WC use	1.12	2.36	3.93	5.25	5.95
Y_9	Bed mobility	1.01	1.26	2.44	4.10	5.74
Y_{10}	Eating	1.00	1.03	1.32	2.52	5.24

Table 3 Estimated mean values of the ADL items, conditional on latent states, for the fixed effect model with $k = 5$.

beginning of the evaluation process. This requirement implies that the aforementioned correlation should be negligible.

The way the correlation between NH effects is represented differs in the two LM models. In the random effect model, the single parameter ρ specifies it, whereas in the fixed effect model we have one correlation coefficient for every pair of elements sharing the same position in the vectors β_1 and γ_1 . In our five-state models, we have $\hat{\rho} = -0.118$ with a standard error equal to 0.221, while the maximum estimated absolute correlation between elements of $\hat{\beta}_1$ and $\hat{\gamma}_1$ is lower than 0.05. Similar results are obtained for other values of k . Therefore, we can conclude that in the LTCF data, NH effects on the initial and transition probabilities are uncorrelated, and the proposed ranking procedures can be safely adopted.

In Table 4, the two rankings for the fixed and random effect LM models with $k = 5$, together with the values of \hat{p}_h and i_h , are reported. This table shows that for both models the i_h indicator is generally more spread on the 0-1 interval than \hat{p}_h . However, the four rankings are quite similar, although some noteworthy discrepancies exist, especially for NHs with a small number of residents (see for instance NH number 30). These are likely due to the shrinkage effect of the random effect model.

To better understand the second ranking procedure, we also depict the caterpillar plot of \hat{p}_h for the fixed effect model with $k = 5$ (see Figure 1). In such a plot, the dots are pinpointed at the values of \hat{p}_h , while the bars represent the associated OIs. NHs are sorted according to the value of \hat{p}_h , from the lowest to the highest. As an illustration, we consider the NH labelled by number 4 in Table 4, whose standardised 180-day ahead transition matrix

NH (h)	n_h	\hat{p}_h -based ranking				i_h -based ranking			
		RE model		FE model		RE model		FE model	
		\hat{p}_h	rank	\hat{p}_h	rank	i_h	rank	i_h	rank
1	96	0.508	41.0	0.445	41.0	0.0	41.0	0.013	41.0
2	8	0.729	32.0	0.708	31.0	0.375	32.0	0.425	30.0
3	21	0.762	16.0	0.774	17.0	0.588	15.5	0.550	16.0
4	7	0.724	34.0	0.693	33.0	0.312	34.0	0.412	33.0
5	73	0.735	28.0	0.739	26.0	0.425	29.5	0.438	26.5
6	23	0.763	14.0	0.783	13.0	0.588	15.5	0.588	13.0
7	80	0.758	19.0	0.768	20.0	0.550	20.0	0.525	20.5
8	89	0.756	22.0	0.763	21.0	0.525	22.0	0.512	22.5
9	38	0.733	30.0	0.729	29.0	0.425	29.5	0.425	30.0
10	45	0.612	40.0	0.526	39.0	0.062	40.0	0.075	39.0
11	49	0.792	4.0	0.803	8.0	0.788	4.5	0.725	7.5
12	10	0.725	33.0	0.663	34.0	0.325	33.0	0.262	34.0
13	63	0.741	26.0	0.735	27.0	0.450	26.0	0.438	26.5
14	10	0.680	36.0	0.574	36.0	0.100	36.0	0.112	36.0
15	17	0.758	21.0	0.770	18.0	0.550	20.0	0.550	16.0
16	19	0.792	5.0	0.823	5.0	0.788	4.5	0.788	5.0
17	36	0.767	13.0	0.775	15.0	0.650	13.0	0.550	16.0
18	39	0.736	27.0	0.727	30.0	0.438	27.0	0.425	30.0
19	72	0.758	20.0	0.761	22.0	0.550	20.0	0.512	22.5
20	9	0.775	11.0	0.806	7.0	0.675	11.0	0.712	9.0
21	64	0.735	29.0	0.732	28.0	0.425	29.5	0.425	30.0
22	12	0.812	2.0	0.854	3.0	0.888	2.0	0.912	1.5
23	21	0.745	24.0	0.741	25.0	0.475	24.5	0.475	25.0
24	17	0.781	8.0	0.801	9.0	0.725	7.5	0.725	7.5
25	8	0.732	31.0	0.703	32.0	0.425	29.5	0.425	30.0
26	8	0.797	3.0	0.860	2.0	0.850	3.0	0.900	3.0
27	29	0.790	6.0	0.813	6.0	0.750	6.0	0.775	6.0
28	39	0.762	15.0	0.770	19.0	0.588	15.5	0.538	19.0
29	19	0.772	12.0	0.793	10.0	0.662	12.0	0.662	10.5
30	5	0.777	9.0	0.883	1.0	0.712	9.0	0.875	4.0
31	12	0.664	37.0	0.555	37.0	0.088	37.5	0.100	37.0
32	15	0.746	23.0	0.754	23.0	0.488	23.0	0.525	20.5
33	27	0.776	10.0	0.790	11.0	0.688	10.0	0.662	10.5
34	45	0.783	7.0	0.789	12.0	0.725	7.5	0.650	12.0
35	20	0.652	38.0	0.494	40.0	0.088	37.5	0.062	40.0
36	18	0.761	17.0	0.780	14.0	0.588	15.5	0.550	16.0
37	26	0.635	39.0	0.548	38.0	0.075	39.0	0.088	38.0
38	20	0.745	25.0	0.743	24.0	0.475	24.5	0.488	24.0
39	39	0.825	1.0	0.848	4.0	0.938	1.0	0.912	1.5
40	15	0.760	18.0	0.774	16.0	0.575	18.0	0.550	16.0
41	15	0.690	35.0	0.634	35.0	0.112	35.0	0.162	35.0

Table 4 Indices \hat{p}_h and i_h and respective rankings for the fixed effect (FE) and the random effect (RE) models with $k = 5$. The sample size n_h of each NH is also included.

according to the fixed effect model is

$$\hat{\mathbf{H}}_4 = \begin{pmatrix} 0.872 & 0.126 & 0.002 & 0.000 & 0.000 & 0.000 \\ 0.037 & 0.675 & 0.275 & 0.012 & 0.001 & 0.000 \\ 0.001 & 0.062 & 0.604 & 0.303 & 0.028 & 0.002 \\ 0.000 & 0.003 & 0.075 & 0.488 & 0.386 & 0.048 \\ 0.000 & 0.000 & 0.007 & 0.100 & 0.539 & 0.354 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}.$$

For this model, it is straightforward to compute $\hat{p}_4 = 0.693$ from the matrix above. This means that in NH 4 there is an overall probability of 0.693 that a resident will not incur a worsening of his/her physical limitations in a 180-day-long period of time. Also, in the caterpillar plot of Figure 1 two dashed lines are drawn to extend the bounds of the OI of this NH. In this way, it is easier to identify the NHs with a significantly better or worse performance. Specifically, there are $B_4 = 12$ better-performing NHs and $W_4 = 5$ worse-performing NHs, so that $i_4 = 0.412$. Interestingly, Figure 1 shows that the growth rate of \hat{p}_h is rather high for the ten worst-performing NHs, while it considerably decreases for the subsequent NHs. The same dynamic seems to occur with respect to the width of the OIs, with the exception of the best and the worst NH. Similar patterns are observed in the caterpillar plot derived from the random effect model with the same number of classes (not shown).

In order to evaluate the appropriateness of the proposed ranking procedures, it is important to bear in mind that the i_h -based ranking relies on the assumption of normality for \hat{p}_h ; see Section 3.4. In this regard, we argue that \hat{p}_h could also be interpreted as the estimated proportion of patients not worsening their physical health status, had all $n = 1292$ patients been treated by the h -th NH. As such, the normal approximation for its sampling distribution seems plausible given the sample size at hand. In any case, 95% OIs are always included in the 0-1 interval, since none of the estimated \hat{p}_h is too close to the boundaries. This is also the case for 95% confidence intervals, which are always wider than the associated OIs.

To check the robustness of our ranking procedure to a further extent, we have built the same set of rankings as in Table 4 for some akin models, that is, for models with a similar number of latent states. Specifically, we considered $k = 4$ and $k = 6$. For these models, results in terms of covariate effects as well as of interpretability of the latent states are very close to those of the models with $k = 5$. The Spearman correlation coefficient of any pair of rankings obtained with $k = 4, 5, 6$ is never lower than 0.95. This finding denotes a remarkable level of robustness across the model specification (fixed versus random effects), the index defining the ranking (\hat{p}_h versus i_h), and the number of latent states ($k = 4, 5, 6$).

Finally, it is worth to remark that NH effects could also be measured on the linear predictor scale (*i.e.*, by $\hat{\gamma}_1$ for the fixed effect model and by $\hat{\sigma}_s E(S_h | \mathbf{Y}_h)$ for the random effect model) rather than on the probability scale. Because of the parametrisation of Equations (3) and (5), the rankings produced by these parameters are identical to those defined by the respective performance

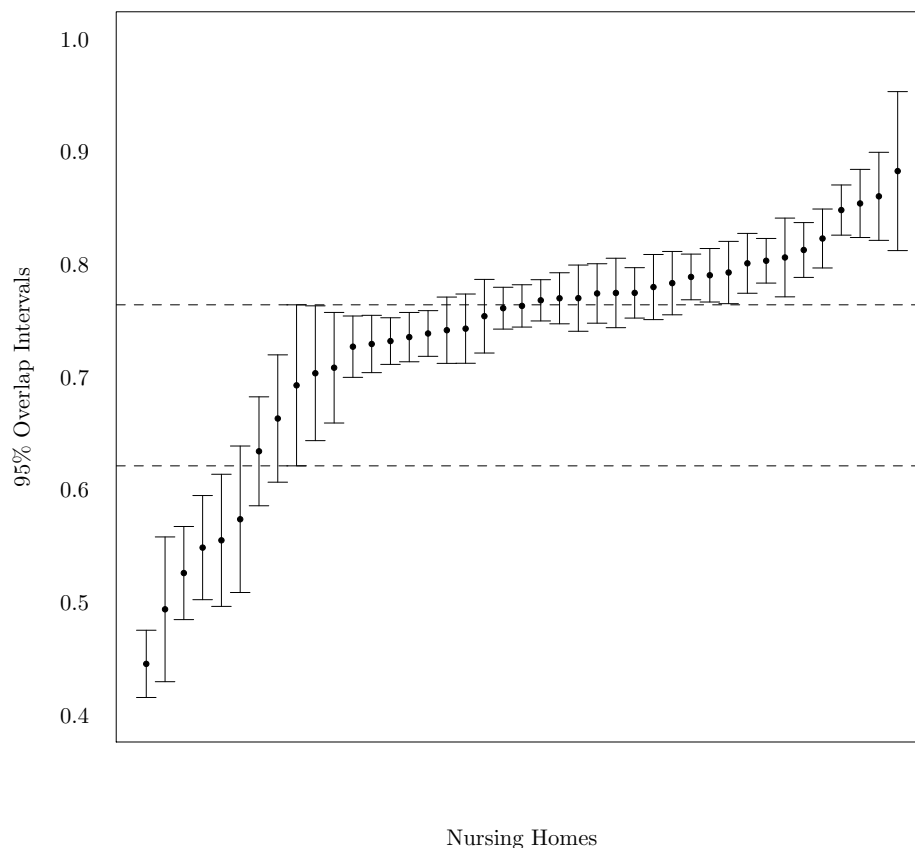


Fig. 1 Caterpillar plot of \hat{p}_h for the fixed-effect model with $k = 5$. NHs are ordered according to the value of \hat{p}_h (dots).

indicators \hat{p}_h , whereas discrepancies may arise in the rankings based on the i_h index. For the sake of completeness, we have built also these alternative i_h -based rankings, but no substantially different results occur. However, we argue that the \hat{p}_h indicator has an undoubtedly clearer interpretation than the effects on the linear predictor scale have.

5 Conclusions

In this paper, we consider the problem of ranking a set of nursing homes (NHs) according to their capability to improve or keep stable over time the physical

limitations status of their residents. Given the latent nature of the trait of interest (physical limitation in the ADLs) and the longitudinal perspective in measuring care quality, we rely on latent Markov (LM) models with individual as well as NH group effects on the unobservable variable. The proposed LM model also accounts for dropout due to resident death. This approach enables us to define an NH performance index based on properly standardised 180 day-ahead transition matrices. Two ranking procedures are then proposed. The first is solely based on the performance index, whereas the second also accounts for uncertainty due to its estimation, in a multiple comparison perspective. An application to a longitudinal dataset coming from the *Long Term Care Facilities* (LTCF) Programme of Regione Umbria (Italy) is carried out. This dataset contains information on the health status of elderly residents in the Umbrian NHs.

Two competing LM models are analysed. Specifically, NH effects are modelled as fixed or random effects, respectively. Random effect models are generally preferable since they provide more reliable estimates for NHs for which just a few observations are available. Conversely, estimation of fixed effect LM models might fail or be highly unstable in these settings. In such a case, small structures with few residents have to be discarded and, as a consequence, they cannot be ranked. This kind of problem occurs for the LTCF data examined in this paper. Moreover, random effect models are characterised by a lower estimation variability for relevant parameters. However, the computational burden is typically greater for random effect models. Furthermore, compared to fixed effect models, they involve an additional assumption regarding the distribution of the random effects. Here, a bivariate normal distribution for the NH random effects is assumed. Such an assumption seems suitable in this context, since NH effects can be reasonably viewed as an average of several different components. In this respect, note that further developments might be also pursued extending the proposal of Bartolucci et al (2014) to the multilevel case.

A sensitivity analysis is performed to determine the appropriate number of latent states of the LM models in use. Although for both models five latent states are finally selected, the rankings deriving from models with four and six latent states are also considered. Thus, a robustness check of our rankings can be performed with respect to three factors: model specification, number of latent states and type of ranking. The Spearman rank correlation of any pair of rankings obtained in this way was never lower than 0.95.

Though the overall robustness of our rankings is promising, it is worth to remark that even with a high level of correlation between two rankings there might be sensible differences in the ranks for some NHs. Therefore, like for any other statistical tool, results from these procedures may help inform policy makers owning subject matter knowledge, but they cannot just be blindly applied for decisional purposes.

Finally, we recall that our rankings are based on a one-dimensional latent trait, while health status is a multidimensional phenomenon. Clearly, similar rankings can be built with respect to different health domains, one at a time.

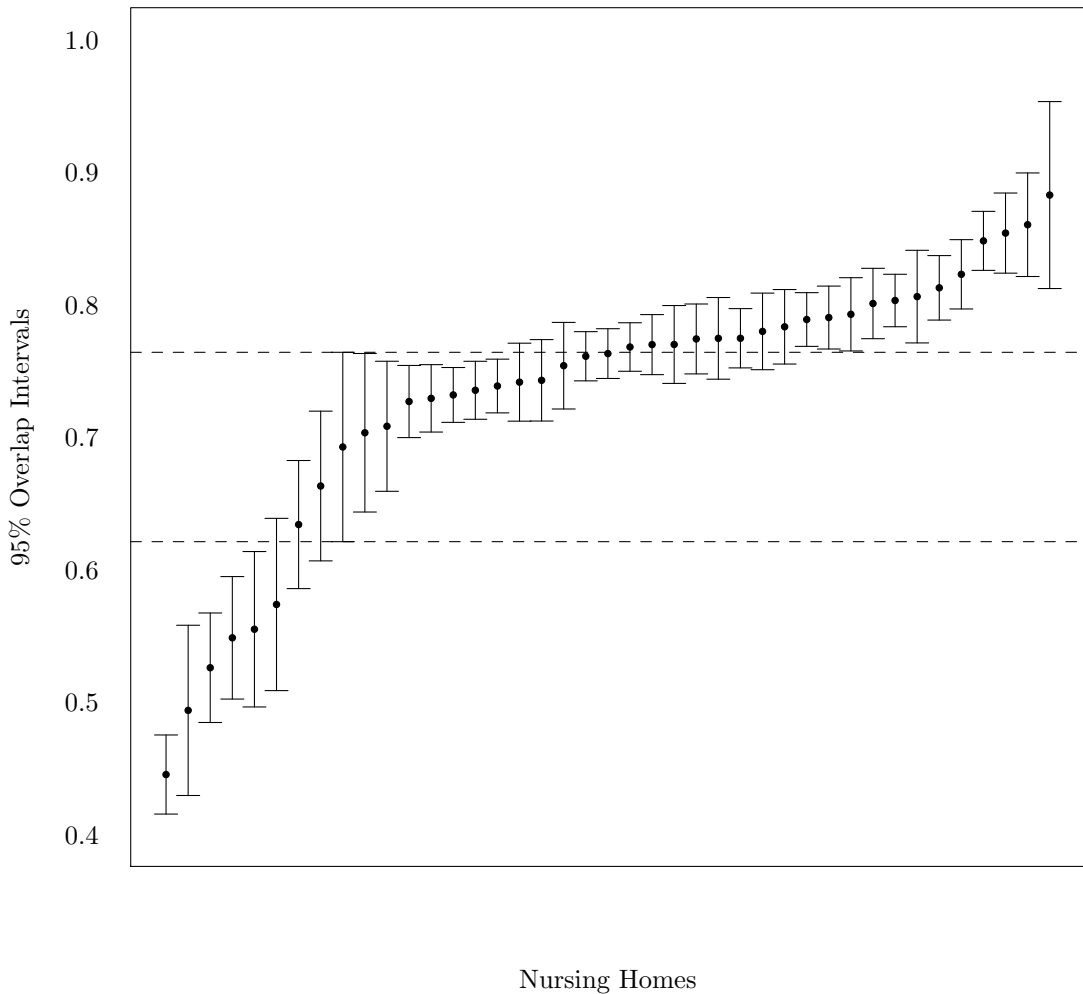
In this respect, further research is worthy to investigate how to obtain an overall ranking from those related to the different health domains at hand.

References

- Afshartous D, Preston RA (2010) Confidence intervals for dependent data: Equating non-overlap with statistical significance. *Computational Statistics & Data Analysis* 54(10):2296–2305
- Altman RM (2007) Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association* 102(477):201–210
- Arling G, Kane RL, Lewis T, Mueller C (2005) Future development of nursing home quality indicators. *The Gerontologist* 45(2):147–156
- Arling G, Lewis T, Kane RL, Mueller C, Flood S (2007) Improving quality assessment through multilevel modeling: The case of nursing home compare. *Health Services Research* 42(31):1177–1199
- Bacci S, Pandolfi S, Pennoni F (2014) A comparison of some criteria for states selection in the latent markov model for longitudinal data. *Advances in Data Analysis and Classification* 8:125–145
- Bartolucci F, Lupparelli M, Montanari GE (2009) Latent Markov model for longitudinal binary data: an application to the performance evaluation of nursing homes. *The Annals of Applied Statistics* 3(2):611–636
- Bartolucci F, Pennoni F, Vittadini G (2011) Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics* 36(4):491–522
- Bartolucci F, Farcomeni A, Pennoni F (2013) *Latent Markov Models for Longitudinal Data*. *Statistics in the Social and Behavioural Sciences*, Chapman & Hall/CRC
- Bartolucci F, Bacci S, Pennoni F (2014) Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63(2):267–288
- Bartolucci F, Pandolfi S, Pennoni F (2017) Lmest: An r package for latent markov models for longitudinal categorical data. *Journal of Statistical Software* 81(4):1–38
- Castle NG, Ferguson JC (2010) What is nursing home quality and how is it measured? *The Gerontologist* 50(4):426–442
- Colombi R, Martini G, Vittadini G (2017) Determinants of transient and persistent hospital efficiency: The case of Italy. *Health Economics* 26(S2):5–22
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39(1):1–38
- Fletcher R (1987) *Practical methods of optimization*, 2nd edn. New York: John Wiley & Sons
- Gnaldi M, Ranalli MG (2010) Composite indicators of scientific research: the robustness of university rankings based on composite measures. In: *Proceed-*

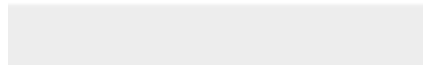
- ings of the XLV Scientific Meeting of the Italian Statistical Society. June, 16-18, Padua
- Gnaldi M, Ranalli MG (2016) Measuring university performance by means of composite indicators: A robustness analysis of the composite measure used for the benchmark of Italian universities. *Social Indicators Research* 129:659–675
- Goldstein H, Healy MJR (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A* 158(1):175–177
- Grilli L, Rampichini C (2009) Multilevel models for the evaluation of educational institutions: a review. In: Monari P, Bini M, Piccolo D, Salmaso L (eds) *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, Physica-Verlag HD, Heidelberg, pp 61–80
- Hirdes JP, Ljunggren G, Morris JN, Frijters DH, Finne Soveri H, Gray L, Björkgren M, Gilgen R (2008) Reliability of the interRAI suite of assessment instruments: a 12-country study of an integrated health information system. *BMC Health Services Research* 8(1):277
- Kim H, Jung YI, Sung M, Lee JY, Yoon JY, Yoon JL (2015) Reliability of the interRAI Long Term Care Facilities (LTCF) and interRAI Home Care (HC). *Geriatrics & Gerontology International* 15(2):220–228
- Kitagawa EM (1964) Standardized comparisons in population research. *Demography* 1(1):296–315
- Koukounari A, Moustaki I, Grassly NC, Blake IM, Basáñez MG, Gambhir M, Mabey DC, Bailey RL, Burton MJ, Solomon AW (2013) Using a nonparametric multilevel latent Markov model to evaluate diagnostics for trachoma. *American Journal of Epidemiology* 177(9):913–922
- Makai P, Brouwer WB, Koopmanschap MA, Stolk EA, Nieboer AP (2014) Quality of life instruments for economic evaluations in health and social care for older people: a systematic review. *Social Science & Medicine* 102:83–93
- Marozzi M, Bolzan M (2016a) An index of household accessibility to basic services: A study of Italian regions. *Social Indicators Research* URL <https://doi.org/10.1007/s11205-016-1440-0>
- Marozzi M, Bolzan M (2016b) Skills and training requirements of municipal directors: a statistical assessment. *Quality and Quantity* 50:1093–1115
- Maruotti A (2011) Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review* 79(3):427–454
- Maruotti A, Rocci R (2012) A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine* 31(9):871–886
- Montanari GE, Pandolfi S (2018) Evaluation of long-term health care services through a latent Markov model with covariates. *Statistical Methods & Applications* 27(1):151–173
- Montanari GE, Doretto M, Bartolucci F (2017a) A multilevel latent Markov model for the evaluation of nursing homes' performance, mPRA Working Paper. Available at <http://mpra.ub.uni-muenchen.de/80691/>
- Montanari GE, Doretto M, Bartolucci F (2017b) An ordinal Latent Markov model for the evaluation of health care services. In: *SIS 2017 Statistics and*

- Data Science: new challenges, new generations, pp 707–712
- Montanari GE, Doretto M, Bartolucci F (2017c) Statistical assessment of public health care services: a multilevel Latent Markov model. In: Proceedings of the 8th Scientific Conference on Innovation & Society, Statistical Methods for Evaluation and Quality. September, 6th-7th, 2017, Naples
- Oakes D (1999) Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)* 61:479–482
- Pennoni F, Vittadini G (2013) Two competing models for ordinal longitudinal data with time-varying latent effects: an application to evaluate hospital efficiency. *Journal of Methodological and Applied Statistics* 15:53–68
- Phillips CD, Zimmerman D, Bernabei R, Jonsson PV (1997) Using the resident assessment instrument for quality enhancement in nursing homes. *Age and Ageing* 26(S2):77–81
- Pohle J, Langrock R, van Beest FM, Schmidt NM (2017) Selecting the number of states in hidden Markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics* 22(3):270–293
- Rampichini C, Grilli L, Petrucci A (2004) Analysis of university course evaluations: from descriptive measures to multilevel models. *Statistical Methods and Applications* 13(3):357–373
- Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464
- Vermunt JK, Langeheine R, Bockenholt U (1999) Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics* 24(2):179–207
- Vittadini G, Minotti SC (2005) A methodology for measuring the relative effectiveness of healthcare services. *IMA Journal of Management Mathematics* 16(3):239–254
- Wiggins LM (1973) Panel analysis: Latent probability models for attitude and behavior processes. Jossey-Bass
- Zucchini W, MacDonald IL (2009) Hidden Markov models for time series, 1st edn. Chapman & Hall/CRC





Click here to access/download
Attachment to Manuscript
FixedVsRandLM.R2.tex






Click here to access/download
Attachment to Manuscript
FixedVsRandLM.R2.bbl



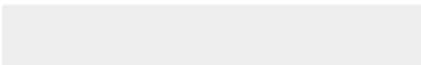



Click here to access/download
Attachment to Manuscript
BiblioSIR.bib



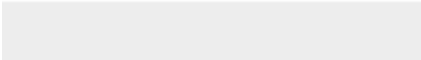



Click here to access/download
Attachment to Manuscript
example.eps





Click here to access/download
Attachment to Manuscript
spmpsi bst





Click here to access/download
Attachment to Manuscript
spphys.bst



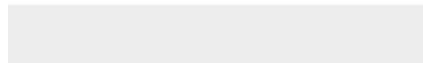


Click here to access/download
Attachment to Manuscript
svjour3.cls





Click here to access/download
Attachment to Manuscript
svglov3.clo





Click here to access/download
Attachment to Manuscript
spbasic.bst

