Received: 12 October 2020 | Revised: 9 July 2021 | Accepted: 13 July 2021

DOI: 10.1002/jae.2882

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

The dynamic interdependence in the demand of primary and emergency secondary care: A hidden Markov approach

Mauro Laudicella[1] | **Paolo Li Donni**[1,2]

[1]DaCHE-Danish Center for Health Economics, University of Southern Denmark, Odense, Denmark

[2]Università degli Studi di Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, Palermo, Italy

Correspondence
Paolo Li Donni, Università degli Studi di Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, Viale delle Scienze, Palermo, 90129, Italy.
Email: paolo.lidonni@unipa.it

Summary

This paper develops an extension of the class of finite mixture models for longitudinal count data to the bivariate case by using a hidden Markov chain approach. The model allows for disentangling unobservable time-varying heterogeneity from the dynamic effect of utilisation of primary and secondary care and measuring their potential substitution effect. Three points of supports adequately describe the distribution of the latent states suggesting the existence of three profiles of low, medium and high users who shows persistency in their behaviour, but not permanence as some switch to their neighbour's profile.

1 | INTRODUCTION

The demand for emergency secondary care (ESC) is growing fast in many publicly funded health care systems absorbing a large share of their resources and jeopardising their financial sustainability (Berchet, 2015). ESC includes visits to the hospital emergency department (ED) and unplanned hospital admissions; it is associated with higher costs and poorer health outcomes than planned secondary care and primary care (PC) for similar health conditions, thus resulting in inefficient allocation of resources and suboptimal health outcomes for the population.

Redirecting the demand of care from ESC to PC has been a long-standing policy objective and the rapid spread of the COVID-19 is adding urgency to implementation of new policies (Busse et al., 2019). In England, for instance, two health policies have been the subject of numerous evaluation studies: the Quality and Outcome Framework introduced financial incentives to improve disease management in PC preventing use of ESC for patients with chronic conditions (Dusheiko et al., 2011; Oxholm et al., 2018); the Equitable Access to Primary Medical Care increased accessibility of PC services by extending opening hours and opening new walk-in centres in selected areas of the country (Dolton & Pathania, 2016; Pinchbeck, 2019). Similar policies have been implemented in other countries in the EU (Iezzi et al., 2014; OECD/WHO, 2019).

PC can substitute ESC for a set of ambulatory care sensitive conditions that can be treated directly in the general physician's (GP) practice, including circulatory and heart conditions such as hypertension and congestive heart failure (Oster & Bindman, 2003). PC can prevent the use of ESC by monitoring and managing patients with chronic conditions, thus reducing the risk of an acute episode, or by monitoring and managing individuals surviving an acute episode, such a stroke or a heart attack, thus reducing the risk of a second episode. Finally, early detection of the onset of disease in PC and timely referral to planned secondary care can reduce the risk of using more expensive and less effective ESC (Starfield et al., 2005). In many other cases, however, PC and ESC respond to different health care needs and treat conditions that

require different levels of specialisation leaving no room for a substitution effect. For instance, individuals experiencing independent health shocks over time or affected by health conditions that increase in severity over time may require access to both PC and ESC. In these cases, PC and ESC can be considered as complementary services that concur in the recovery of full health.[1]

Estimating a substitution effect between PC and ESC requires some considerations on the characteristics of the data generating process. First, PC and ESC are likely to be jointly determined, making the measurement of a substitution effect difficult to achieve in cross-sectional analysis. PC utilisation is likely to affect ESC and vice versa, hence introducing circularity in the identification of a substitution effect. Moreover, unobserved heterogeneity in the individual health status and preferences may affect the demand of both PC and ESC producing an additional source of endogeneity. To deal with these issues, existing empirical applications make use of the exogenous variation in the supply of PC induced by a policy reform over time and across geographical areas. Then, they measure the substitution effect at the level of the GP practice or small geographical area by using a linear panel data model (Dolton & Pathania, 2016; Dusheiko et al., 2011; Fortney et al., 2005; Pinchbeck, 2019; Whittaker et al., 2016) or a Poisson model (Iezzi et al., 2014; Lippi Bruni et al., 2016) and a difference in differences approach, which is corroborated by a propensity score matching or an instrumental variable approach in some applications. These studies find evidence of a substitution effect at the GP practice level, although the patient level substitution effect might be different due to ecological fallacy, and the effect in the total population might be different from the effect in areas targeted by the policy. To overcome these limitations, Atella and Deb (2008)propose a recursive nonlinear model consisting of a system of three simultaneous equations modelling utilisation of PC and public and private specialist care. The model identification is based on the assumption that PC utilisation precedes specialist care, which seems reasonable as patients normally need a GP referral to access public specialist care, but it might be not applicable to model utilisation of ESC as patients can attend the hospital ED directly and without a GP referral.

A second set of considerations apply to the dynamic characteristics of the data generating process. In longitudinal data, PC and ESC are likely to be dynamically related through the patient health conditions. On the one hand, individuals with a poor health endowment who do not see their GP for a long time may increase their risk of developing a severe undetected illness or an acute episode from a pre-existing chronic condition; these individuals are likely to use ESC once their symptoms become acute. For these individuals, using more PC in the past could reduce the risk of using ESC in the present. On the other hand, individuals with a better health endowment may produce a similar history of few GP visits and yet bear a low risk of experiencing an acute episode, and thus using ESC. Increasing PC utilisation in past periods would have little effect on ESC utilisation in the current period for these individuals. Therefore, in order to identify the dynamic relationship between PC and ESC utilisation, we need to be able to disentangle the contribution of past use to current use from the contribution of unobservable heterogeneity in individual's health. Moreover, the data generating process described here is consistent with the modelling of an *intertemporal* substitution effect with past PC utilisation potentially affecting present ESC.

Finally, the dynamic relationship between PC and ESC reflects the idea that individual health status is time varying and persistent, that is, individual health follows a time trajectory in which past health is a good predictor of present health (Contoyannis et al., 2004). Thus, we may expect the unconditional utilisation of health care to be persistent as shown in a growing number of empirical investigations (see Monheit, 2003; Kohn & Liu, 2013; French & Jones, 2004). However, persistency in the context of the dynamic relationship between primary and secondary care has not received a similar level of attention.

This paper contributes to the applied econometric literature by introducing an econometric model that allows for estimating the dynamic relationship between PC and ESC utilisation and their intertemporal substitution effect. The model extends the class of finite mixture models for longitudinal count data (Bago d'Uva, 2005; Deb & Trivedi, 1997) to the bivariate case, in which the number of PC and ESC visits are jointly modelled and depend on a latent process following a first-order Markov chain. This feature allows the researcher to disentangle the contribution of past utilisation and time-varying unobserved heterogeneity on present utilisation. In the economic literature, hidden Markov models (also known as *Markov regime-switching*) have been used to model macroeconomic and financial time series. They have also been applied to study health care utilisation by using count data models consisting of a single equation (Hyppolite & Trivedi, 2012; Alfò & Maruotti, 2010). Our proposed model extends this approach to the bivariate case allowing for time-varying unobserved heterogeneity, state dependence and residual time-specific correlation between PC and ESC.

[1]A detailed discussion on the determinants of health care utilisation can be found in Deb and Trivedi (1997).

We demonstrate an application of the model by estimating the substitution effect between PC and ESC in individuals experiencing an health shock to their circulatory system in the Danish National Health System.

The paper is organised as follows: the next section reviews the relevant econometric literature; section 2 illustrates the model; section 3 describes the data and institutional background for the empirical application; section 4 describes our findings. Conclusions and discussion are in section 5.

2 | PROPOSED MODEL

Longitudinal latent class (finite mixture) regressions are not new in the literature of univariate count data model (Bago d'Uva, 2005). Generally, they are used to model unobserved heterogeneity by assuming that individual effects are approximated using a discrete distribution. This is an alternative representation of heterogeneity, where individuals are drawn from a finite number of latent classes, that can be regarded as types or groups representing individual unobserved characteristics. For example, in the case of primary and secondary health care utilisation these characteristics may refer to unmeasured health status or preference for utilisation.

When the focus is on how current realisations of a count variable depend on its past realisations, a standard strategy within the finite mixture framework is to introduce dynamics in the latent process by assuming that it follows a first-order Markov chain (see, e.g., Hyppolite & Trivedi, 2012; Alfò & Maruotti, 2010). Latent classes translate into a hidden Markov series of states that depend on previous period states. This property allows for serial dependence between observations collected in successive periods, making this approach a useful tool in modelling time-varying unobserved heterogeneity as individuals are allowed to change state at different time occasions. This assumption is consistent with evidence of persistence in individual health status and utilisation of health care (Contoyannis et al., 2004; French & Jones, 2004; Kohn & Liu, 2013; Monheit, 2003).

Let Y_{jit} denote a random count variable measuring health care utilisation, and y_{jit} the observed count j for subject i at time occasion t, with $j = 1, 2$, $i = 1, \dots, n$ and $t = 1, \dots, T$. Let $\boldsymbol{y}_{it}$ be a vector collecting the observed count variable y_{1it} and y_{2it} for each individual in each time period, and denote with $\boldsymbol{y}_i$ the corresponding $T \times 2$ matrix defined by the following subvectors $(\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \dots, \boldsymbol{y}_{iT})$. Let also indicate with $\boldsymbol{\alpha}_{it}$ a vector of time-varying random effects collecting α_{1it} and α_{2it} and with $\boldsymbol{x}_{jit}$ a vector of exogenous covariates of dimension $1 \times h_j$. The proposed model relies on two main assumptions: (*i*) the subject-specific random parameters follow a first-order Markov chain with latent states ξ_c, for $c = 1, \dots, k$,[2] and (*ii*) the latent states make y_{1it} and y_{2it} conditionally independent given the full set of observable explanatory variables, including the past realisations y_{1it-1} and y_{2it-1}. The latter assumption is a form of local independence, since the outcome variables are conditionally independent given the latent process $\boldsymbol{\alpha}_{i1}, \dots, \boldsymbol{\alpha}_{iT}$. From this perspective, latent states aim to capture the relevant time-varying unobserved heterogeneity conditional on the full set of explanatory variables. Thus, different states ξ_c, $c = 1, \dots, k$, are meant to capture different individual attitudes towards health care utilisation.

To model the marginal joint distribution of y_{1it} and y_{2it}, we employ a Bivariate Poisson (BPo) distribution (Lakshminarayana et al., 1999; Famoye, 2010):

$$\boldsymbol{Y}_{it} | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1} \sim \text{BPo}(\mu_{1it}, \mu_{2it}, \lambda) \tag{1}$$

where the conditional distribution of observing y_{1it} and y_{2it} is given as a product of two Poisson marginals with a multiplicative factor λ such that

$$p(\boldsymbol{y}_{it} | \boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) = \frac{\mu_{1it}^{y_{1it}} \mu_{2it}^{y_{2it}} e^{-\mu_{1it}-\mu_{2it}} [1 + \lambda(e^{-y_{1it}} - e^{-q\mu_{1it}})(e^{-y_{2it}} - e^{-q\mu_{2it}})]}{(y_{1it}! y_{2it}!)} \tag{2}$$

with $q = 1 - e^{-1}$ and $\mu_{jit} = \exp\left(\sum_{c=1}^{k} \alpha_{jit}(\xi_c) d_{jit}(\xi_c) + \boldsymbol{x}_{jit} \boldsymbol{\beta}_j + \gamma_{j1} y_{1i,t-1} + \gamma_{j2} y_{2i,t-1} \right)$, with $j = 1, 2$ and $d_{jit}(\xi_c)$ denotes a dummy variable defining whether the ith unit belongs to the latent state ξ_c of the Markov chain at time t. Notice that if $\lambda = 0$, the BPo in (2) reduces to the product of two Poisson with mean μ_{jit}. The joint mean and dispersion matrix of Y_{1it} and Y_{2it} can then be written in block form:

[2]Assuming a discrete, rather than a continuous latent process, avoids parametric assumptions on the structure of the unobserved heterogeneity (e.g., normality), hence providing more flexibility to the model (see Heckman & Singer, 1984).

$$E[\boldsymbol{Y}_{it}] = \begin{bmatrix} \mu_{1it} \\ \mu_{2it} \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \mu_{1it} & \lambda\mu_{1it}\mu_{2it}q^2e^{-q(\mu_{1it}+\mu_{2it})} \\ \lambda\mu_{1it}\mu_{2it}q^2e^{-q(\mu_{1it}+\mu_{2it})} & \mu_{2it} \end{bmatrix}. \tag{3}$$

Correlation coefficient is then given by $\rho_{it} = \lambda\sqrt{\mu_{1it}\mu_{2it}}q^2e^{-q(\mu_{1it}+\mu_{2it})}$. Therefore $\rho = nT^{-1}\sum_i\sum_t\rho_{it}$ can then be used to summarise the correlation between counts. This correlation can be positive, zero, or negative depending on the value of the multiplicative factor parameter λ. However, ρ_{it} depends also on individual characteristics, and thus, it can vary widely as compared to the scalar value λ, which directly captures the time-specific correlation between y_{1it} and y_{2it} conditional on $\boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}$ and the unobservable time-varying individual heterogeneity captured by $\boldsymbol{\alpha}_{it}$.

Therefore, the model allows for three possible sources of correlation between the unobservable determinants of the demand for primary and secondary care. The first is related to state dependence, which occurs when utilisation of PC and ESC in past periods predicts utilisation in the present period. This type of correlation is captured by including a lag of the outcome variables in the equations of the intensity parameter, that is, y_{1it-1} and y_{2it-1} in μ_{1it} and μ_{2it}, thus modelling state dependence within and between $y_{\cdot it}$. State dependence *within outcomes* provides information on the extent to which utilisation of a given type of care depends on its utilisation in the past, for example, individuals who used ESC in the past period may (may not) use it again in the current period if their health condition has not (has) improved. State dependence *between outcomes* can be used to explain utilisation-patterns between different types of care over time. In particular, if lagged states of PC utilisation have a reducing effect on ESC utilisation in the current state, then this could interpreted as a substitution effect between PC and ESC.

Second, correlation between observations over time is captured by individual-specific effects $\boldsymbol{\alpha}_{it}$. This vector of parameters, one for each latent state ξ_c and counts j, captures time-varying heterogeneity in the individuals' propensity to use both PC and ESC. In other words, $\boldsymbol{\alpha}_{it}$ captures the persistence of health care utilisation that is due to individual unobserved heterogeneity. For instance, individual's unmeasured health conditions or preferences for health may gradually strengthen or weaken in the long term following a treatment or a health shock, thus affecting her utilisation of PC and ESC permanently.

Third, the parameter λ, which determines the average conditional residual correlation (ρ), controls for *time-specific* shocks that may affect both outcomes in any given period and are not captured by $\boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}$, or by the subject-specific parameters $\boldsymbol{\alpha}_{it}$. Such correlated period shocks can be the result of an unexpected health deterioration that is occasional and influences utilisation of both PC and ESC in a specific time-period t'. For instance, individuals on a course of medication that is no longer effective may experience a time specific shock that triggers both PC and ESC utilisation before starting a new treatment.

The distribution of the random parameter vectors $\boldsymbol{\alpha}_{it}$ is assumed to follow a first-order homogenous time Markov chain, that is, for each subject i the random parameter vectors $\boldsymbol{\alpha}_{it}$, $t = 1, \dots, T$, follow a first-order Markov chain with states ξ_c, for $c = 1, \dots, k$. Hence, individuals are assumed to be drawn from a discrete latent distribution with k classes in each time period. These classes represent different unobserved states in which individuals are likely to be observed with probability $p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c)$ in state c at time t. Since serial correlation in the unobserved factors is taken into account by assuming a first order Markov chain, the probability mass function of $\boldsymbol{\alpha}_i$ for the two outcomes and for a configuration of latent states $\boldsymbol{\xi}$ can be expressed as follows:

$$p(\boldsymbol{\alpha}_i = \boldsymbol{\xi}) = p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c)\prod_{t=2}^{T} p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}}|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c) \tag{4}$$

where $\boldsymbol{\xi}$ denotes a matrix collecting a realisation of the latent states $\boldsymbol{\xi}_c$ for the two outcomes in each time period and $c, \bar{c} = 1, \dots, k$. Support points $\boldsymbol{\xi}_c$ capture the unobserved time-varying heterogeneity in a flexible non-parametric way, such that the distribution of $\boldsymbol{\alpha}_i$ given in Equation (4) is described by both the initial $p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c)$ and transition probabilities $p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}}|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c)$, which are collected in matrix Π.

In order to take explicitly into account the initial condition problem (Heckman, 1981a), we allowed the initial probabilities $p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c)$ in Equation (4) depends not only on $\boldsymbol{y}_{i0}$, but also on the explanatory variables. This is achieved by using a multinomial logit parameterisation:

$$p\left(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c|\boldsymbol{x}_{i0}, \boldsymbol{y}_{i0}\right) = \frac{\exp(\alpha_c + \boldsymbol{\beta}_c\boldsymbol{x}_{i0} + \boldsymbol{\gamma}_c\boldsymbol{y}_{i0})}{1 + \sum_{c=2}^{k}\exp(\alpha_c + \boldsymbol{\beta}_c\boldsymbol{x}_{i0} + \boldsymbol{\gamma}_c\boldsymbol{y}_{i0})}, \tag{5}$$

with $c = 2, \ldots, k$. Estimated coefficients in (5) are particularly interesting, as they capture how individual conditions prior to the health shocks affect different trajectories of primary and secondary health care utilisation in the subsequent T periods. Notice that the $k-1$ logit parameters in (5) do not impose any parametric restriction on the distribution of $\boldsymbol{\alpha}_i$.

In summary, the set of equations in (2), (4) and (5) defines our Hidden Markov Bivariate Poisson (HMBPo) model.

2.1 | Likelihood inference

The marginal log-likelihood for the proposed model can be obtained as follows:

$$\ell(\boldsymbol{\theta}) = \sum_{i}^{N} \log\left[p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i0})\right] \tag{6}$$

where $\boldsymbol{\theta}$ be the vector collecting all the non redundant model parameters corresponding to the vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$; λ, and the off-diagonal elements of the matrix Π. Given the structure of the model described above the marginal density $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i0})$ is equals to

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i0}) = \sum_{\boldsymbol{\alpha}_{i1}} \cdots \sum_{\boldsymbol{\alpha}_{iT}} \left[p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c|\boldsymbol{x}_{i0}, \boldsymbol{y}_{i0}) \prod_{t=2}^{T} p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}}|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c) \times \times \prod_{t=1}^{T} p(\boldsymbol{y}_{it}|\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1}) \right] \tag{7}$$

which implies the sum over all the possible configurations of $\boldsymbol{\alpha}_{i1}$ given the latent Markov process (MacDonald & Zucchini, 1997) while $p(\boldsymbol{y}_{it}|\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})$ refers to the BPo described in Equation (2). Probability $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{y}_{i0})$ can be efficiently computed by using a forward recursion based on the Baum–Welch algorithm, which is a special case of the EM (Dempster et al., 1977) algorithm used to find the unknown parameters of a latent Markov model. The EM algorithm alternates two steps until convergence. In the first step, named E-step, the conditional expected value of the complete data log-likelihood is computed given the observed data and the corresponding current estimate $\widetilde{\boldsymbol{\theta}}$. Finally the M-step maximises the preceding expected value with respect to $\boldsymbol{\theta}$.

To estimate the model, it is convenient to write Equation (7) in terms of complete log-likelihood:

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) = \sum_{i}^{N} \sum_{c}^{k} \Bigg\{ \sum_{t}^{T} w_{itc} \log[p(\boldsymbol{y}_{it}|\boldsymbol{\alpha}_{it}, \boldsymbol{x}_{it}, \boldsymbol{y}_{i,t-1})] + w_{i1c} \log[p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c|\boldsymbol{x}_{i0}, \boldsymbol{y}_{i0})] + \\ + \sum_{\bar{c}}^{k} z_{ic\bar{c}} \log[p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}}|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c)] \Bigg\} \end{aligned} \tag{8}$$

where w_{itc} denote a dummy variable equal to 1 if subject i is in latent state c at occasion t (i.e., $\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c$) and to 0 otherwise, while z_{icd} counts the number of times subject i changes state from c to $\bar{c}$. In the E-step the conditional expected value of $\ell^*(\boldsymbol{\theta})$ is evaluated by substituting variables w_{itc} and z_{icd} with their corresponding expected values $\tilde{w}_{itc}$ and $\tilde{z}_{icd}$ evaluated at $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$. Efficient computation of these expected probabilities can be implemented using the matrix notation described in Bartolucci (2006) and Bartolucci and Farcomeni (2009). At the M-step the conditional complete log-likelihood $\ell^*\left(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}\right)$ can be maximised by splitting equation in (7) in three separate components.

$$\ell_1^*\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda|\widetilde{\boldsymbol{\theta}}\right) = \sum_{i}^{N} \sum_{c}^{k} \sum_{t}^{T} \tilde{w}_{itc} \log\left[\frac{\mu_{1it}^{y_{1it}} \mu_{2it}^{y_{2it}} e^{-\mu_{1it}-\mu_{2it}}}{(y_{1it}! y_{2it}!)} \times \right. \tag{9}$$

$$\left. \times \left(1 + \lambda(e^{-y_{1it}} - e^{-q\mu_{1it}})(e^{-y_{2it}} - e^{-q\mu_{2it}})\right) \right]$$

$$\ell_2^*\left(\boldsymbol{\beta}|\widetilde{\boldsymbol{\theta}}\right) = \sum_{i}^{N} \sum_{c}^{k} \tilde{w}_{i1c} \log[p(\boldsymbol{\alpha}_{i1} = \boldsymbol{\xi}_c|\boldsymbol{x}_{i0}, \boldsymbol{y}_{i0})] \tag{10}$$

$$\ell_3^*\left(\boldsymbol{\beta}|\widetilde{\boldsymbol{\theta}}\right) = \sum_{i}^{N} \sum_{c}^{k} \sum_{\bar{c}}^{k} \tilde{z}_{ic\bar{c}} \log[p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}}|\boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c)] \tag{11}$$

The first component ℓ_1^* can be maximised using a standard iterative algorithm of Newton–Raphson type for count data models. Similarly for ℓ_2^*, a modified version of Newton–Raphson described in Colombi and Forcina (2001) has been used. Finally, an explicit solution is available to maximise ℓ_3^*, which consists of letting each $p(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_{\bar{c}} | \boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\xi}_c)$ proportional to $\sum_i \tilde{z}_{ic\bar{c}}$ for $c, \bar{c} = 1, \ldots, k$ (see, e.g., Bartolucci & Farcomeni, 2009).

Finally, we follow McLachlan and Peel (2004, Chap. 2). to derive the information matrix and compute the observed information matrix $\boldsymbol{J}(\boldsymbol{\theta})$—denoted as minus the numerical derivative of the score vector $\boldsymbol{s}(\boldsymbol{\theta})$, which corresponds to the first derivative of $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The full set of estimation function has been implemented in a series of Mata routines for StataCorp. v.2014 available upon request from the author. Appendix SA provides more details on EM algorithm.

The choice of an appropriate number of latent states is based on the Bayesian Information Criterion which penalise for the number of parameters in the model (see McLachlan & Peel, 2004). The model with the lowest BIC is usually preferred. The number of parameters is given by the sum of the following: (*i*) the number of rows in the vectors: $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\gamma}_2$; (*ii*) $2 \times k$ α's intercepts for the intensity parameters μ in Equation (3) and $k-1$ intercepts α_c in Equation (5); *iii*) one association parameters λ, and *iv*) $k(k-1)$ of off-diagonal elements of the matrix Π, which corresponds to the number of independent transition probabilities.

A relevant piece of information that can be recovered from the HMBPo estimates is the share ω_c of individuals in the population belonging to the unobserved state c. This is obtained by using the posterior distribution of latent states ξ_c. In particular, for each individual i at time occasion t the posterior probability $p\left(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c | \boldsymbol{x}_{it}, \boldsymbol{y}_{it}\right)$ is computed via Bayes's formula and backward recursion involved by Equation (4). Subsequently the posterior weight ω_c for each latent state is obtained as follows:

$$\omega_c = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{T} \sum_{t=1}^{T} p\left(\boldsymbol{\alpha}_{it} = \boldsymbol{\xi}_c | \boldsymbol{x}_{it-1}, \boldsymbol{y}_{it-1}\right) \right], \text{ with } c = 1, \ldots, k \tag{12}$$

A hypothesis of particular interest is whether the effect of unobserved individual heterogeneity is effectively time-varying, hence supporting the use of the HMBPo model. This hypothesis can be directly tested by checking if the off-diagonal elements of the matrix Π are jointly different from zero, which indicates that the effect of unobserved factors on the PC and ESC is not time-constant. In such a case, conventional models disregarding the dynamic in the unobserved components are not suitable. Moreover, the constraint that all the off-diagonal transition probabilities are equal to zero generates a boundary problem invalidating standard likelihood ratio tests. Therefore, we follow an alternative test proposed by Bartolucci (2006) and specifically designed to deal with this issue in hidden Markov models.

3 | DATA AND INSTITUTIONAL BACKGROUND

We provide an application of the HMBPo model to study the dynamic interdependence between the demand of PC and ESC in the Danish National Health System (DNHS). To this aim, we selected a population of individuals age 50+ who survived an initial health shock to their circulatory system in 2012 and modelled their patterns of utilisation of PC and ESC in the five years following the health shock up to 2017. PC includes physical contacts with the GP for any cause including home visits; ESC includes visits to the hospital ED for any cause including visits resulting in a hospital admission and visits treated at the ED not resulting in an admission. We defined an initial health shock as the first ESC admission for a disease of the circulatory system [3]excluding patients with history of ESC admissions for any cause in the five years before the initial health shock. These selection criteria produce a relatively large and homogeneous population of 9146 patients who are likely to access PC and ESC during the study period, thus producing sufficient variation in the outcome variables and reducing model identification issues due to excess of zeros (Atella & Deb, 2008). Moreover, setting the initial health shock as the starting point of our study allows us to examine a population with relatively similar initial characteristics reducing potential confounding effects due to past utilisation patterns. Finally, we divided time in six months intervals to capture a sufficient amount of variation in the outcome variables in each interval.

Data were extracted from the Danish National Patient Register (NPR), which collects very rich administrative and clinical information on individuals accessing GP PC and hospital secondary care in Denmark. We had access to dates of PC and ESC visits, including up to 20 clinical diagnoses for each visit. All Danish residents are assigned a unique personal identification number that is used to link primary and secondary care visits at the individual level and additional information from other registers. We linked data on date of death from the Register of Causes of Death, individual annual

[3]International Classification of Diseases v10, codes I00-I99 reported as primary diagnosis.

TABLE 1 Summary statistics

Variable name	Mean	SD
ESC visits	0.20	0.64
PC visits	3.70	3.58
female	0.40	—
age	66.08	9.57
living alone	0.32	—
migrant	0.06	—
income	29,328.7	64,738.49
travel time to ESC	14.57	10.75
Charlson index	0.67	0.73
total diagnoses	1.00	1.30
AMI (I21)	0.20	0.40
chronic ischaemic heart disease (I25)	0.02	0.14
pulmonary heart disease (I26)	0.04	0.19
other forms of heart disease (I30-I52)	0.29	0.46
stroke (I62-I63)	0.19	0.39
other cerebrovascular diseases (I6X)	0.04	0.19
diseases of arteries (I7X)	0.03	0.16
other diseases of veins (I8X)	0.05	0.22

income from the Income Statistics Register, individuals' place of residence and living alone status from the Central Person Register (see Thaygesen et al., 2011 for an overview on Danish registries). Hospital addresses are publicly available and were used to calculate travel distances between individuals and closest hospital ED.

The DNHS is a universal health care system and PC and ESC are free of charge at the point of access. This provides ideal conditions for our study as the demand of health care is not confounded by differences in patients' health insurance coverage or ability to pay for services. GPs offer PC services to Danish residents enrolled in their list with 30 per cent of their income provided by a fix capitation payment, and the rest by a fee for service system centrally regulated by the government; GPs are also the gate keepers of specialist care and elective secondary care. All residents in Denmark have to enroll in a GP practice of their choice within a catchment area of their place of residence (OECD, 2017).

ESC services are accessed by calling an emergency number managed by the Emergency Medical Coordination Centres, which assess the urgency of the call and direct the patients to the closest hospital if appropriate. ESC is delivered by joint acute facilities, that is, hospital EDs with the capacity of receiving patients suffering from all kind of health problems (Christiansen & Vrangbaek, 2018);49 of these departments were active during our study period and evenly spread across the country serving a population of about 5.5 millions of Danes. ESC and all secondary care services are reimbursed by the government through a prospective payment system based on DRG tariffs.

Table 1 reports descriptive statistics for our study population of 9146 individuals followed for 5 years in 6 months intervals. We worked with a perfectly balanced panel of 91,460 patient-period observations with an average of 3.70 PC visits and 0.20 ESC visits per patient-period. This large difference is expected as PC is normally the first point of contact for treating a large range of non-acute health conditions, while ESC visits address less prevalent acute conditions. Individual travel time to the closest ED is on average 15 min with a large standard deviation of about 10 minutes, which makes ESC more accessible to some individuals than others. The average age of our study population is 66, 40% are females, 32% live alone, 6% are migrant, annual income is about 29,000 euros per person. In terms of health conditions, 20% had an initial health shock due to a AMI, 29% due to other forms of heart diseases, 19% due to a stroke, while the remaining population had a shock due to other cerebrovascular diseases, and other diseases of arteries and veins. Finally, individuals have on average one diagnosis and a Charlson comorbidity index of 0.67 points.

Overall observed correlation between PC and ESC utilisation is about 14 per cent suggesting some degree of interdependence similarly to other studies (e.g., Riphahn et al., 2003). Figure 1 shows the distribution of PC visits at growing number of ESC visits suggesting the two variables follow a similar pattern. PC visits appear increasingly dispersed at higher level of ESC utilisation suggesting the possible existence of different groups of individuals with different propensity towards health care utilisation, for example, two groups of frequent users of ESC with different attitudes towards PC.

Table 2 illustrates the dynamic relationship between PC and ESC. It shows variation in present utilisation over time in individuals who used PC or ESC in the past period. Consuming ESC in the previous period is associated with consuming slightly less than global average ESC in current period, and two times more than global average PC. In contrast, consuming

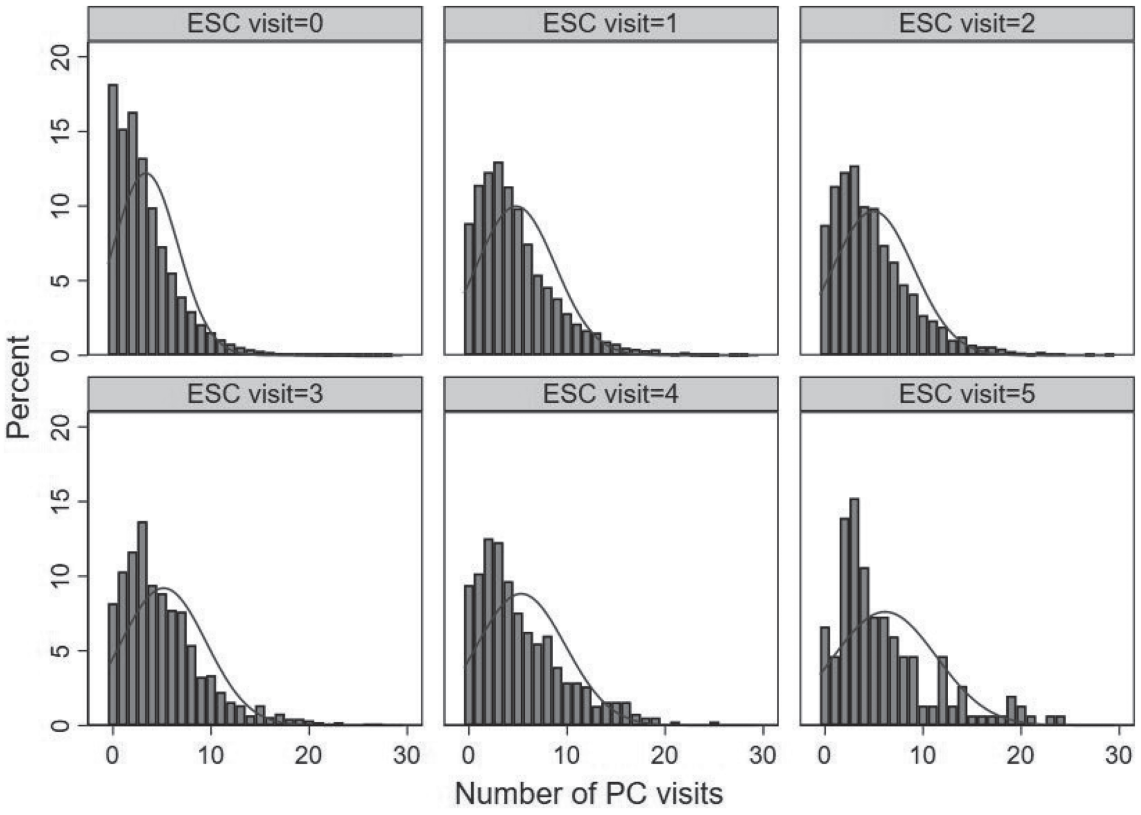

FIGURE 1 Observed distribution of primary care (PC) visits by increasing number of emergency secondary care (ESC) visits

	$PC_{t-1} > 0$			$ESC_{t-1} > 0$		
Time period	**♯ of Obs.**	**Mean(ESC)**	**Mean(PC)**	**♯ of Obs.**	**Mean(ESC)**	**Mean(PC)**
2	8852	0.19	6.99	2382	0.30	7.73
3	8542	0.18	6.56	1033	0.31	7.89
4	8394	0.17	6.27	935	0.33	7.53
5	8336	0.17	6.32	899	0.34	7.51
6	8325	0.17	6.22	915	0.33	7.58
7	8238	0.16	6.38	934	0.33	7.76
8	8293	0.17	6.45	893	0.39	8.12
9	8285	0.18	6.34	921	0.39	8.05
10	8210	0.23	6.48	912	0.50	8.19

TABLE 2 Dynamic distribution of the observed health care utilisation

PC in the previous period is associated with consuming up to two times more than global average of both ESC and PC in current period (PC and ESC global averages are reported in Table 1).

4 | RESULTS

We estimated the HMBPo on our study population and compared it against two alternative models in order to assess performance in fitting the data and predicting the relationship between PC and ESC. The first model consists of two separate Poisson equations with random effects (REPo) that allow for individual time-fixed heterogeneity by using a dynamic panel specification with normally distributed random effects, but model PC and ESC utilisation as two independent processes. The second is a Bivariate Poisson model (BPo) that allows for the interdependence between PC and ESC, but disregards individual unobserved heterogeneity and the panel structure of the data. Both REPo and BPo include controls for initial conditions (Wooldridge, 2005).[4]

[4]Initial condition controls y_{1i0}, y_{2i0} and average of time-varying x for each equation of PC and ESC.

TABLE 3 Model selection criteria

Model	k	♯	Log-lik.	BIC	AIC
REPo		78	−214171.56	429054.57	428499.12
BPo		77	−230744.65	462191.63	461643.31
HMBPo	1	69	−232546.85	465723.05	465231.69
	2	99	−216997.86	434898.7	434193.72
	3	131	−212765.01	426724.88	425792.02
	4	165	−212732.54	426970.06	425795.08

Table 3 reports the AIC and BIC scores for the three models. [5] With respect to the HMBPo model, AIC and BIC suggest that three latent states seem sufficient to capture time-varying individual heterogeneity in the HMBPo model.

4.1 | The relationship between types of care

Table 4 reports estimated coefficients for state dependence and substitution effects and socioeconomic variables, which can be interpreted as semielasticities. Complete estimates including time dummies and individual health indicators are reported in Appendix SB; estimates of HMBPo initial conditions are in Appendix SC.

All the three models predict a positive state dependence in PC utilisation: an additional PC visit in the past period results in a 2.6%–2.8% increment in present PC visits according to the REPo and the HMBPo models, and 8.7% increment according to the BPo model. Notice that HMBPo and BPo estimates are conditional to the ESC distribution, while REPo are unconditional. In contrast, REPo predicts a negative state dependence in the utilisation of ESC with an additional ESC visit in the past resulting in −9.5% visits in the present, while BPo and HMBPo predicts an increment of 16.4% and 12.7%, respectively. Evidence of positive state dependence in utilisation of health care is found by several studies based on survey data (Hyppolite & Trivedi, 2012; Contoyannis et al., 2004; Kohn & Liu, 2013).

Large differences between the three models emerge in the estimates of the substitution effects. With respect to the effect of PC on ESC, REPo and BPo predict that an additional PC visit in the previous period increases ESC visits by 0.9% and 0.4%, respectively, although the latter is not statistically different from zero. Namely, these models suggest that consumption of PC is a weak complement or independent to consumption of ESC. In contrast, the HMBPo model suggests a substitution effect of PC to ESC with an additional PC visit in past period reducing ESC visits by 12.6%. Evidence of a similar substitution effect is found by several studies based on exogenous variation from new policies (Dolton & Pathania, 2016; Fortney et al., 2005; Pinchbeck, 2019; Whittaker et al., 2016; Lippi Bruni et al., 2016). With respect to the effect of ESC on PC, an additional ESC visit in the previous period increases PC visits by 1.2% according to the REPo model again suggesting complementarity. In contrast, it reduces by −1.2% according to the BPo and by a −5.0% according to the HMBPo suggesting a substitution effect.

To put these estimates into context, a variation of one PC visit is well within its standard deviation of 3.58 visits, while a variation of one ESC visit is considerably larger than its standard deviation of 0.64 visits. Therefore, HMBPo estimates suggest a large scope for reducing ESC by increasing PC visits, while the reverse has a limited scope and policy interest. Similarly, state dependence in the utilisation of ESC is small in magnitude as compared with state dependence in PC.

Finally, ignoring unobserved individual heterogeneity results in spurious inference on the residual correlation between PC and ESC, which is determined by model parameter λ. According to the BPo model, a time-specific shock results in a residual variation of PC and ESC in the same direction, while the HMBPo model predicts a residual variation with opposite directions, that is, a residual increment (reduction) of ESC and reduction (increment) of PC.

Finally, the HMBPo model allows us to expand the analysis of the substitution effect by examining the dynamic response of PC and ESC visits over time after individuals experience their initial health shock. To this end, we interacted the cross-effects y_{1it-1} and y_{2it-1} with time and calculated their average marginal effects assuming the first time period after the health shock as the baseline. The results are plotted in Figure 2. With respect to the demand for ESC, the substitution

[5]The HMBPo and the BPo are not nested as they solve the initial condition problem following two different approaches. The HMBPo solves the initial conditions by explicitly specifying a separate Equation (5) (see e.g. Heckman, 1981a, 1981b), while the BPo includes controls for y_{1i0}, y_{2i0} and average of time-varying x in addition to x_{ijt}, $y_{1i,t-1}$ and $y_{2i,t-1}$ following Wooldridge (2005). When $k = 1$, Equation (5) is not estimated, since individuals are univocally assigned to one state from the first period. We assessed model fitting by comparing the difference between observed and predicted ESC and PC visits estimated by the RE and the HMBPo (with $k = 3$). We reported observed and predicted marginal probabilities obtained using the two models in Table S3 in Appendix SB. The HMBPo provides the smallest difference between observed and predicted values in more than 95% of the distribution of PC and ESC.

TABLE 4 Estimated coefficients

	REPo model		BPo model		HMBPo model	
	PC visit	ESC visit	PC visit	ESC visit	PC visit	ESC visit
constant[a]	0.2766***	−3.2624***	0.2874***	−2.6367***	0.7169***	−2.367***
	(0.07)	(0.18)	(0.02)	(0.11)	−0.07	−0.15
female	0.0857***	−0.0245	0.0519***	−0.0078	0.0086	−0.0653***
	(0.01)	(0.03)	(0.00)	(0.02)	(0.01)	(0.02)
age 50–59	0.0882	0.2009	0.0893***	0.2502**	0.0654	0.2280
	(0.06)	(0.17)	(0.02)	(0.11)	(0.06)	(0.14)
age 60–69	0.2253***	0.1444	0.1957***	0.2142**	0.0924	0.0841
	(0.06)	(0.17)	(0.02)	(0.11)	(0.06)	(0.14)
age 70–79	0.3125***	0.2839*	0.2536***	0.3052***	0.1429**	0.1647
	(0.06)	(0.17)	(0.02)	(0.11)	(0.06)	(0.14)
age 80–89	0.2584***	0.5135***	0.2296***	0.4512***	0.1431**	0.3786***
	(0.06)	(0.18)	(0.02)	(0.11)	(0.06)	(0.14)
age 90+	0.0025	0.8098***	0.0545	0.6176***	0.0432	0.7884**
	(0.09)	(0.24)	(0.03)	(0.13)	(0.18)	(0.31)
single	−0.0854***	0.1500***	−0.0445***	0.1328***	−0.0346***	0.1666***
	(0.01)	(0.04)	(0.00)	(0.02)	(0.01)	(0.02)
migrant	0.0634**	0.1547**	0.0534***	0.1337***	−0.0411**	0.0223
	(0.03)	(0.07)	(0.01)	(0.03)	(0.02)	(0.04)
inc. quartile 2	0.0273	−0.0902**	0.0192***	−0.0932***	0.0079	−0.1228***
	(0.02)	(0.05)	(0.01)	(0.02)	(0.01)	(0.03)
inc. quartile 3	−0.0017	−0.0575	−0.0035	−0.0844***	−0.0438***	−0.1481***
	(0.02)	(0.05)	(0.01)	(0.02)	(0.01)	(0.03)
inc. quartile 4	−0.0880***	−0.0481	−0.0743***	−0.0877***	−0.1214***	−0.1197***
	(0.02)	(0.05)	(0.01)	(0.03)	(0.01)	(0.03)
trav. time	0.0043***	−0.0129***	0.0026***	−0.0109***	0.0015***	−0.0149***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
trav. time^2	−0.0001***	0.0001**	−0.0001***	0.0001***	−0.0001	0.0001***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
PC	0.0258***	0.0085***	0.0872***	0.0041	0.0277***	−0.1206***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ESC	0.0119***	−0.0945***	−0.0119***	0.1636***	−0.0496***	0.1271***
	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
λ			1.0269***		−1.5761***	
			(0.03)		(0.11)	

Note: Standard errors are in parentheses. Each model includes controls for primary diagnosis, Charlson index, number of comorbidities and time dummies.

[†] Average of the support points based on the posterior probabilities in the HMBPo.

[*]Significant at 10%.

[**]Significant at 5%.

[***]Significant at 1%.

effect of PC reaches its peak six months after the initial health shock and gradually decreases thereafter (right panel of Figure 2). Similarly, the effect of using ESC in the previous period is larger six months after the health shock (i.e., time period 2); thereafter the relationship changes sign becoming negative with respect to the baseline period. Turning to the demand of PC, an additional PC visit in the previous period has a small incremental effect that remains relatively constant over time after the initial health shock (left panel of Figure 2). In contrast, the substitution effect of an ESC visit increases one year after the initial health shock (i.e., time period 3) and stays relatively constant thereafter.

4.2 | Unobserved heterogeneity and subject-specific parameters

The HMBPo model provides a tool to study persistence in utilisation of care due to individual unobserved heterogeneity allowing for time-varying heterogeneity.

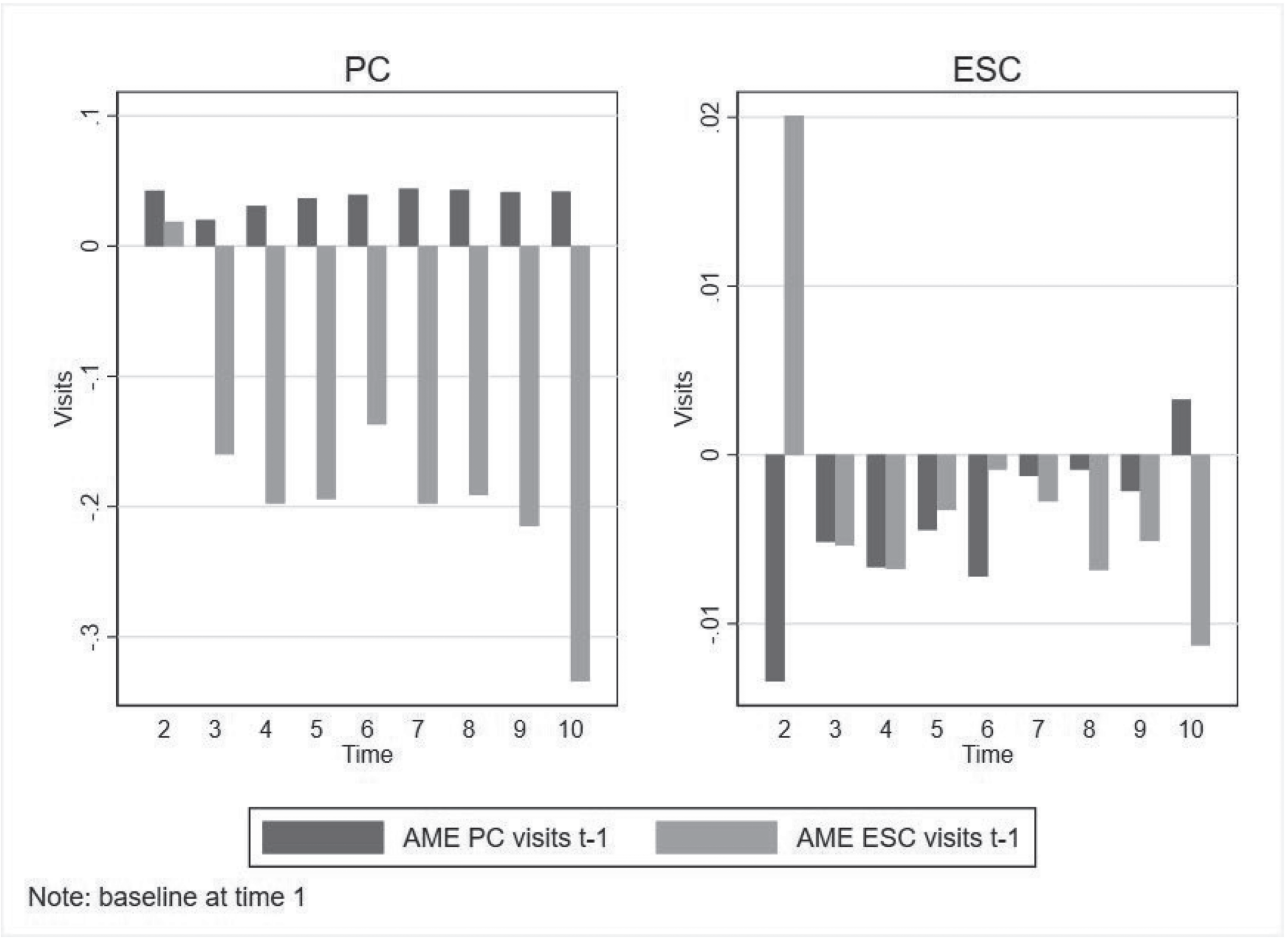

FIGURE 2 Estimated average marginal effects of emergency department (ED) and general physician (GP) visits

TABLE 5 Estimated support points ξ_c for the subject-specific parameters

Latent	PC visit		ESC visit	
State	$\alpha_1(\xi_c)$	SD	$\alpha_2(\xi_c)$	SD
1	−0.4072***	−0.01	−3.9335***	0.17
2	0.9360***	0.04	−2.0819***	−0.12
3	1.7291***	−0.07	−0.8990***	−0.15

Note: Standard errors are in parentheses.

*Significant at 10%.

**Significant at 5%.

***Significant at 1%.

Table 5 reports HMBPo estimates of support points for three unobserved types (classes) of users who differ in their propensity to use PC and ESC.

In order to characterise the latent classes of individuals in Table 5, we predict their probability of using a given number of PC and ESC visits. Table 6 shows that individuals in the first class are more likely to make up to one PC visit (79% probability—pp) and zero ESC visits (97 pp), individuals in the second class are more likely to make two to three PC visits (41 pp) and zero ESC visits (86 pp), while individuals in the third class are more likely to make more than four PC visits (90 pp) and zero to one ESC visits (89 pp) in every period. Although making one or more ESC visit has a lower probability than making none in all latent groups, this probability increases exponentially from 3 pp in the first class, to 13 pp in the second class and 34 pp in the third class. Hence, these latent classes suggest the existence of three types of individuals with increasing propensity to use PC and ESC, that is, 'lower' (Class 1), 'medium' (Class 2) and 'heavy' users (Class 3). Such a classification is illustrated in Figure 3 showing the joint distribution of PC and ESC conditional on latent types.

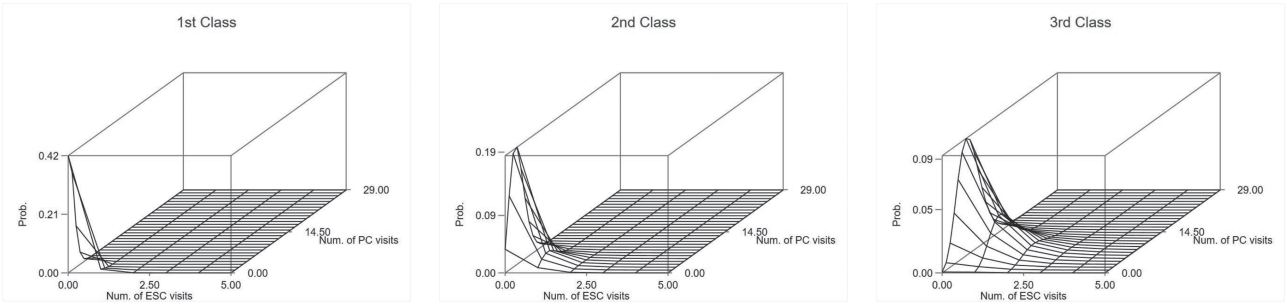

FIGURE 3 Predicted joint distribution of PC and ESC visits by latent types

TABLE 6 Observed and predicted probabilities of health care utilisation by latent classes

	PC visits				ESC visits			
	Obs.	1st Class	2nd Class	3rd Class	Obs.	1st Class	2nd Class	3rd Class
0	0.1720	0.4349	0.0466	0.0015	0.8961	0.9742	0.8653	0.6643
1	0.1475	0.3560	0.1365	0.0093	0.0595	0.0241	0.1124	0.2333
2	0.1589	0.1512	0.2048	0.0284	0.0271	0.0012	0.0150	0.0624
3	0.1320	0.0448	0.2107	0.0592	0.0108	0.0003	0.0035	0.0190
4	0.0999	0.0105	0.1676	0.0943	0.0047	0.0001	0.0013	0.0077
5 or 5+	0.2898	0.0026	0.2338	0.8073	0.0018	0.0001	0.0024	0.0134

The HMBPo model can also be used to examine how time-fixed and time-varying individual heterogeneity influence persistence in utilisation of health care over time. Differently from REPo, HMBPo allows individuals to change class membership over time moving from low users to high users and vice versa. A change of class could occur, for instance, after a health shock or a treatment that leave a permanent mark on individual unmeasured health and preferences. Such changes are summarised by the matrix of transition probabilities:

$$\hat{\Pi} = \begin{bmatrix} 0.80 & 0.18 & 0.02 \\ 0.10 & 0.84 & 0.06 \\ 0.01 & 0.16 & 0.83 \end{bmatrix}. \tag{13}$$

The transition matrix shows that a low user (first row) is very likely to remain in the same class over the time of the study with a probability of 80% (first row, first column), while her probability of becoming a medium user is 18% (first row, second column), and her probability to become a heavy user is very low with only 2%(first row, last column). A similar information is reported for individuals who are medium users (second row), who are very likely to stay in their class (84 pp) or join low users (10 pp), and high users (third row), who are likely to remain in their class (83 pp) or join the medium users (16 pp). Thus, the transition matrix suggests a high degree of persistency in individual heterogeneity, but it also predicts that individuals may change to their neighbour class with a probability of about 17%.

Figure 4 reports class membership probabilities in every time occasion. The probability of joining the class of low users increases over time up to period 6 and then become stable, while the same path with a decreasing direction is followed by the probability of joining the class of medium users. In contrast, the probability of belonging to the class of high users seems quite stable over time, suggesting that new entries and exits from this class balances out. Period 6 marks two and a half years after the initial health shock experienced by individuals in our study, during this period some individuals are likely to have followed a treatment and gradually recovered, thus changing their profile from medium to low users. In contrast, entries and exits from the high user class seem to balance out in every period and thus may have a weaker link with the events that follows the initial health shock.

Finally from Equation (12) of the HMBPo, it is possible to calculate the share of the study population that belongs to each class: low users are 28% of the population, while medium and high users are 53% and 19% respectively.

We performed a formal test of the hypothesis that individual heterogeneity is time invariant by estimating the HMBPo model under the hypothesis that the transition matrix is diagonal, that is, absence of time-varying heterogeneity. The asymptotic distribution of an LR test statistic when the parameters are on the boundary of the parameter space has been studied extensively in the literature (for a review see Silvapulle & Sen, 2004) Bartolucci (2006) demonstrates that the likelihood ratio statistic has null asymptotic distribution of $\bar{\chi}^2$ type. This testing procedure requires to draw a num-

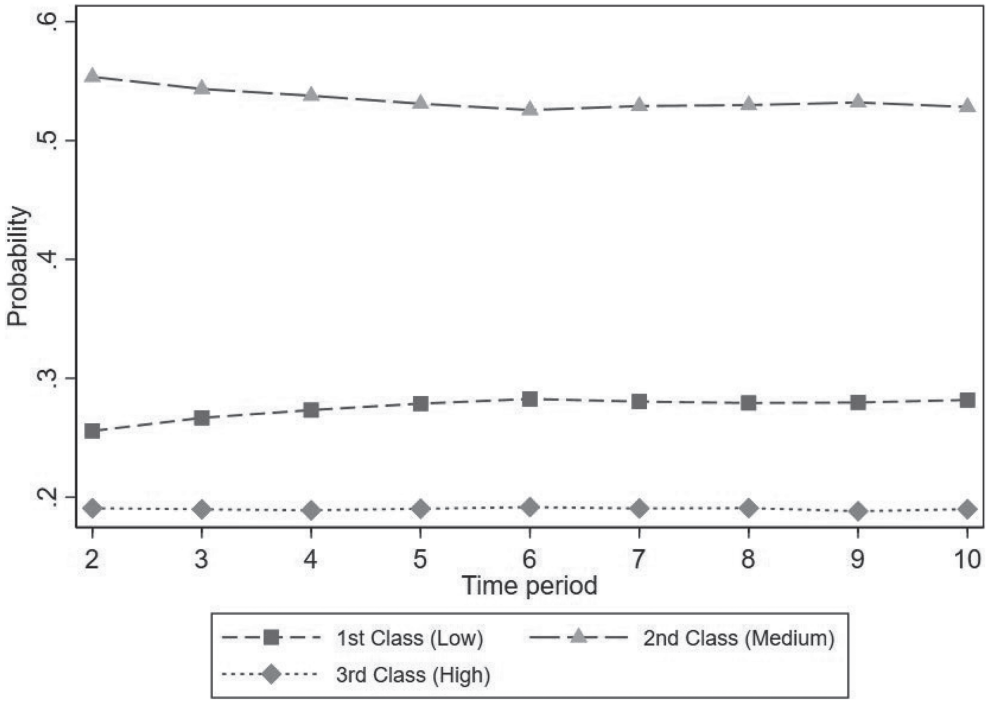

FIGURE 4 Estimated average probabilities of latent states in each period

bers of parameter vectors from the asymptotic distribution of the unconstrained maximum likelihood estimator and then computing the proportion of vectors that violate the hypothesis of absence of transition between latent states. The log-likelihood for the restricted model is equal to −214225.92, that compared to that one of the unrestricted model reported in Table 3 leads to a likelihood ratio statistic with 6 dof rejecting the null hypothesis of time-constat heterogeneity with a p value $<10^{-3}$.

5 | CONCLUSION

This paper introduced a new econometric model to study the dynamic interdependence between the demand of PC and ESC and to measure their potential substitution effect. The model extends the class of finite mixture models for longitudinal count data to the bivariate case by using a hidden Markov chain approach. The model captures some of the distinctive characteristics of the demand of care studied in the literature (Contoyannis et al., 2004; Deb & Trivedi, 1997). Utilisation of PC and ESC are allowed to be jointly determined as patients may access both type of care in the same time occasion. State dependence within and across outcomes is captured by including a lag of PC and ESC in the two outcome equations reflecting the fact that past utilisation is often a strong predictor of present. Persistency in individual health and preferences is modelled by individual specific parameters following a discrete latent variable and capturing unobservable heterogeneity. Assuming a first-order Markov chain in the distribution of the latent allows for disentangling unobservable time-varying heterogeneity from the dynamic effect of utilisation of care. Finally, time-specific non-permanent shocks that may affect both outcomes are captured by a residual correlation parameter.

We provided an empirical application of the model to study the demand of PC and ESC in patients experiencing a shock to their circulatory system in the Danish National Health System. We find evidence of a substitution effect with an additional PC visit in the previous period reducing present ESC visits by about 12.6 percentage points suggesting a fair scope for intervention to policies aiming at redirecting part of the demand of care from ESC to PC. This is in line with results from other empirical studies identifying the substitution effect from the exogenous variation of a policy reform increasing the capacity and accessibility of PC (Pinchbeck, 2019; Lippi Bruni et al., 2016). The proposed model extends previous studies removing the need of an exogenous source of variation for identifying the model, thus allowing the researcher to test for a substitution effect in any segment of the demand of PC and ESC. This could be particularly important from a policy perspective as it allows for exploring the potential scope of a new policy before the latter is implemented in a pilot or countrywide.

Finally, we find evidence that individual profiles of utilisation of health care are persistent, but not permanent. They have a probability of about 17% of changing during the time of our study and move to their closest neighbour (from low

to medium user, from medium to low or to high, and from high to medium), with the majority of switches occurring during the first two and a half years after the initial health shock. These findings contribute to the existing literature on persistency in the demand of health care and consumer behaviour (Hyppolite & Trivedi, 2012; Monheit, 2003; French & Jones, 2004).

A final comment regards possible extensions of the model. In the present application, the distribution of ESC is overdispersed and includes a large number of zeros invalidating inference from a simple Poisson model. The HMBPo provides a partial solution by capturing overdispersion and excess of zeros in the unobserved individual heterogeneity (Deb & Holmes, 2000). However, this approach does not fully address the issue of excess of zeros if the latter are produced by a distinct data generating process. Alternative solutions could be achieved by extending the HMBPo to a bivariate negative binomial model (e.g. Famoye, 2010), which allows for overdispersed counts by estimating an additional parameter, or to a bivariate zero-inflated poisson model (e.g. Liu & Tian, 2015), whichallows for excess of zero counts by using a two-parts model. Finally the version of the HMBPo model presented here was developed for a balanced panel. However, a bias can arise from individuals dropping out before the end of the study (e.g., due to death) and if this event is influenced by their profile of utilisation. There are various ways to account for such an informative dropout. For example, the HMBPo model could be extended allowing the number of time occasions to vary between subjects by adapting the recursions used in the Baum–Welch algorithm to the unbalanced case, while model estimation does not need any relevant adjustment. Alternatively, the researcher can include an additional equation and explicitly model dropout as a time to event outcome based on a subject specific hazard function. Future work could be devoted to extending the model in this direction and addressing the needs of specific applications.

ACKNOWLEDGMENTS

Funded by the EU's Horizon 2020 research and innovation programme under MSCA grant No 832513. Open Access Funding provided by Universita degli Studi di Palermo within the CRUI-CARE Agreement. [Correction added on 18 May 2022, after first online publication: CRUI funding statement has been added.]

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/datasets/laudicella001/].

ORCID

Mauro Laudicella https://orcid.org/0000-0001-5322-4452
Paolo Li Donni https://orcid.org/0000-0001-9074-5089

REFERENCES

Alfò, M., & Maruotti, A. (2010). Two-part regression models for longitudinal zero-inflated count data. *The Canadian Journal of Statistics*, *38*(2), 197–216. https://doi.org/10.1002/cjs.10056

Atella, V., & Deb, P. (2008). Are primary care physicians, public and private sector specialists substitutes or complements? Evidence from a simultaneous equations model for count data. *Journal of Health Economics*, *27*(3), 770–785. https://doi.org/10.1016/j.jhealeco.2007.10.006

Bago d'Uva, T. (2005). Latent class models for use of primary care: Evidence from a british panel. *Health Economics*, *14*(9), 873–892. https://doi.org/10.1002/hec.1047

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *68*(2), 155–178. https://doi.org/10.1111/j.1467-9868.2006.00538.x

Bartolucci, F., & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, *104*(486), 816–831. https://doi.org/10.1198/jasa.2009.0107

Berchet, C. (2015). *Emergency care services: Trends, drivers and interventions to manage the demand*. OECD Health Working Papers 83. OECD Publishing.

Busse, R., Klazinga, N., Panteli, D., & Quentin, W. (2019). *Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies*. World Health Organization. Regional Office for Europe.

Christiansen, T., & Vrangbaek, K. (2018). Hospital centralization and performance in Denmark-Ten years on. *Health Policy*, *122*(4), 321–328. https://doi.org/10.1016/j.healthpol.2017.12.009

Colombi, R., & Forcina, A. (2001). Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, *88*(4), 1007–1019. https://doi.org/10.1093/biomet/88.4.1007

Contoyannis, P., Jones, A., & Rice, N. (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics*, *19*(4), 473–503. https://doi.org/10.1002/jae.755

Deb, P., & Holmes, A. M. (2000). Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models. *Health Economics*, *9*(6), 475–489. https://doi.org/10.1002/1099-1050(200009)9:6<475::AID-HEC544>3.0.CO;2-H

Deb, P., & Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, *12*(3), 313–336. https://doi.org/10.1002/(SICI)1099-1255(199705)12:3<313::AID-JAE440>3.0.CO;2-G

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B: Methodological*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dolton, P., & Pathania, V. (2016). Can increased primary care access reduce demand for emergency care? Evidence from England's 7-day GP opening. *Journal of Health Economics*, *49*, 193–208. https://doi.org/10.1016/j.jhealeco.2016.05.002

Dusheiko, M., Gravelle, H., Martin, S., Rice, N., & Smith, P. C. (2011). Does better disease management in primary care reduce hospital costs? Evidence from english primary care. *Journal of Health Economics*, *30*(5), 919–932. https://doi.org/10.1016/j.jhealeco.2011.08.001

Famoye, F. (2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics*, *37*(6), 969–981. https://doi.org/10.1080/02664760902984618

Fortney, J. C., Steffick, D. E., Burgess, J. F. Jr., Maciejewski, M. L., & Petersen, L. A. (2005). Are primary care services a substitute or complement for specialty and inpatient services? *Health Services Research*, *40*(5 Pt 1), 1422–1442. https://doi.org/10.1111/j.1475-6773.2005.00424.x

French, E., & Jones, J. B. (2004). On the distribution and dynamics of health care costs. *Journal of Applied Econometrics*, *19*(6), 705–721. https://doi.org/10.1002/jae.790

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, *52*(2), 271–320. https://doi.org/10.2307/1911491

Heckman, J. J. (1981a). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in labor markets*, Chapter 3 (pp. 91–140). University of Chicago Press.

Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some monte carlo evidence. In C. F. Manski & D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Chapter 3 (pp. 114–178). MIT Press.

Hyppolite, J., & Trivedi, P. (2012). Alternative approaches for econometric analysis of panel count data using dynamic latent class models (with application to doctor visits data). *Health Economics*, *21*(Suppl. 1), 101–128. https://doi.org/10.1002/hec.2813

Iezzi, E., Lippi Bruni, M., & Ugolini, C. (2014). The role of GP's compensation schemes in diabetes care: Evidence from panel data. *Journal of Health Economics*, *34*, 104–120. https://doi.org/10.1016/j.jhealeco.2014.01.002

Kohn, J. L., & Liu, J. S. (2013). The dynamics of medical care use in the British Household Panel Survey. *Health Economics*, *22*(6), 687–710. https://doi.org/10.1002/hec.2845

Lakshminarayana, J., Pandit, S., & Rao, K. S. (1999). On a bivariate poisson distribution. *Communications in Statistics - Theory and Methods*, *28*(2), 267–276. https://doi.org/10.1080/03610929908832297

Lippi Bruni, M., Mammi, I., & Ugolini, C. (2016). Does the extension of primary care practice opening hours reduce the use of emergency services? *Journal of Health Economics*, *50*, 144–155. https://doi.org/10.1016/j.jhealeco.2016.09.011

Liu, Y., & Tian, G.-L. (2015). Type I multivariate zero-inflated poisson distribution with applications. *Computational Statistics & Data Analysis*, *83*, 200–222. https://doi.org/10.1016/j.csda.2014.10.010

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series* (Vol. *110*). CRC Press.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Monheit, A. C. (2003). Persistence in health expenditures in the short run: Prevalence and consequences. *Medical Care*, *41*(7), III53–III64.

OECD. (2017). *Primary Care in Denmark*. OECD Publishing. https://doi.org/10.1787/9789264269453-en

Oster, A., & Bindman, A. B. (2003). Emergency department visits for ambulatory care sensitive conditions: Insights into preventable hospitalizations. *Medical Care*, *41*(2), 198–207. https://doi.org/10.1097/01.MLR.0000045021.70297.9F

Oxholm, A. S., Kristensen, S. R., & Sutton, M. (2018). Uncertainty about the effort-performance relationship in threshold-based payment schemes. *Journal of Health Economics*, *62*, 69–83. https://doi.org/10.1016/j.jhealeco.2018.09.003

Pinchbeck, E. W. (2019). Convenient primary care and emergency hospital utilisation. *Journal of Health Economics*, *68*, 102242. https://doi.org/10.1016/j.jhealeco.2019.102242

Riphahn, R. T., Wambach, A., & Million, A. (2003). Incentive effects in the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics*, *18*(4), 387–405. https://doi.org/10.1002/jae.680

Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Inequality, order and shape restrictions*. John Wiley & Sons.

Starfield, B., Shi, L., & Macinko, J. (2005). Contribution of primary care to health systems and health. *The Milbank Quarterly*, *83*(3), 457–502. https://doi.org/10.1111/j.1468-0009.2005.00409.x

Whittaker, W., Anselmi, L., Kristensen, S. R., Lau, Y.-S., Bailey, S., Bower, P., Checkland, K., Elvey, R., Rothwell, K., Stokes, J., & Hodgson, D. (2016). Associations between extending access to primary care and emergency department visits: A difference-in-differences analysis. *PLoS Medicine*, *13*(9), e1002113. https://doi.org/10.1371/journal.pmed.1002113

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, *20*(1), 39–54. https://doi.org/10.1002/jae.770

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Laudicella, M., & Li Donni, P. (2022). The dynamic interdependence in the demand of primary and emergency secondary care: A hidden Markov approach. *Journal of Applied Econometrics*, *37*(3), 521-536. https://doi.org/10.1002/jae.2882