EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 19/01
February 2019

A nonlinear dynamic factor model of health and medical treatment

by
Franco Peracchi
(Georgetown University and EIEF)
Claudio Rossetti
(University of Naples Federico II and CSEF)

A nonlinear dynamic factor model of health and medical treatment*

Franco Peracchi
Georgetown University, EIEF and University of Rome Tor Vergata †

Claudio Rossetti
University of Naples Federico II and CSEF ‡

February 5, 2019

Abstract

Quantitative assessments of the relationship between health and medical treatment are of great importance to policy makers. However, simply looking at the raw correlation between health and medical care is unlikely to give the right answer because of endogeneity problems. We overcome these problems by formulating and estimating a tractable dynamic factor model of health and medical treatment where individual observed health outcomes are driven by the individual's latent health stock. The dynamics of latent health reflects both exogenous health depreciation and endogenous health investments. Our model allows us to investigate the effect of medical treatment on current health, as well as on future medical treatment and health outcomes. We estimate the model by maximum simulated likelihood and minimum distance methods using a rich longitudinal data set from Italy obtained by merging a number of administrative archives. These data contain detailed information on medical drug use, hospitalization, and mortality for a representative sample of elderly hypertensive patients. Our findings show that medical care consumption is highly correlated over time, and this relationship depends on both permanent and time-varying observed and unobserved heterogeneity. They also show that medical drug use significantly maintains future health levels and prevents transitions to worse health. These results suggest that policies aimed at increasing the awareness and the compliance of of hypertensive patients help reduces cardiovascular risks and consequent hospitalization and mortality.

Keywords: Dynamic panel data models, latent variable models, factor models, maximum simulated likelihood, minimum distance method, health dynamics, medical treatment, drug consumption, mortality

JEL codes: C15, C33, C35, C81, D12, I12, J14

* We thank Anne Case, Angus Deaton, Tullio Jappelli, Mario Padula and Arthur van Soest for useful discussions. We also thank seminar participants at CSEF and participants at the 6th Italian Congress of Econometrics and Empirical Economics for helpful comments. Franco Peracchi acknowledges financial support from MIUR PRIN 2015FMRE5X.

† E-mail: fp211@georgetown.edu

‡ E-mail: claudio.rossetti@unina.it

1 Introduction

Although medical treatment is expected to have a positive effect on health, in Grossman's original work (Grossman 1972) and in several papers since (see for example Muurinen 1982) the correlation between health status and medical treatment has been found to be negative. This is hardly surprising, as "people tend to seek medical care when they are sick, not when they are well" (Case & Deaton 2005). In fact, simply looking at the raw correlation between health and medical care is unlikely to give the right answer because of endogeneity arising from unobservable determinants of health deterioration (Grossman 2000).

In this paper, we overcome the endogeneity problem by formulating and estimating a tractable dynamic factor model of health and medical treatment. Dynamic factor models have been extensively used in macroeconomics and finance since their introduction by Geweke (1977) and Sargent & Sims (1977) as a way of capturing in a parsimonious way the cross-sectional and dynamic correlations between multiple series. More recent applications in macroeconomics include also non traditional topics such as mortality rates (French & O'Hare 2013). Moreover, there has been growing study in factor models for the kind of high-dimensional data increasingly available in economics and finance, which are characterized by both a large cross-sectional dimension and a large time dimension (see for example Bai & Wang 2014 and Bai & Li 2016. Nevertheless, to our knowledge much less attention has been devoted to dynamic factor models for microeconomic panel data.

Specifically, we propose a dynamic factor model of health and medical treatment in which observable health outcomes are driven by latent individual health, lagged health outcomes, and other observable individual characteristics. We extend the existing literature on the dynamic relationship between health and medical treatment in several ways. First, we model the dynamics of latent health as a function of both exogenous health depreciation and endogenous health investments. This allows us to identify both the direct and the indirect effects – through latent health – of medical treatments on health outcomes, and to investigate the role of medical treatments in the health production function and their effect on future medical use and other health outcomes. Second, by including mortality among the observed health outcomes, we account for endogenous attrition. Third, we introduce nonlinearity in a natural way into our model to accommodate mixed (continuous and discrete) outcomes. Finally, because we allow for both state dependence and latent health dynamics, our model can be viewed either as a generalization of dynamic panel data models or as an extension of state-space models to microeconomic panel data. As a result, our model encompasses as special cases several models frequently found in the literature. Because of its generality, our

model may also be applied to a variety of other settings.

We estimate the model by maximum simulated likelihood and minimum distance methods using the annual longitudinal data employed by Atella et al. (2006). These data have been constructed by combining a number of administrative archives maintained by one regional unit of the Italian national health-care system and provide very detailed micro-level information on medical drug use, hospitalization, and mortality. Because our data contain objective measures of health from administrative health-care records, they offer two important advantages relative to studies based on survey data that only record general self-reported health measures, such as self-rated health on a Likert-type scale. First, they do not present problems of unit and item non-response, or bias effects due to interaction with interviewers, typical of survey data. The chances of measurement errors is also much smaller. Second, by providing detailed information on medical treatments, they make it possible to explore patients' decision-making in relation to specific clinical conditions, and therefore allow one to derive more precise conclusions concerning the relation between medical treatment and health, in particular the role of compliance with medical protocols.

Our sample consists of hypertensive patients born between 1920 and 1939 (aged 61–80 in year 2000) and followed each year from 1997 to 2002. The reason we concentrate on hypertensive patients is because hypertension, by its distinctive nature, provides helpful insights into the relation between medical care and health. In fact, hypertension is a chronic asymptomatic pathology affecting a large share of the adult population, a fraction that actually rises with age. Since hypertension is a chronic pathology, patients should take their medication regularly. However, since it is also asymptomatic, they generally do not feel ill and so may not strictly adhere to medical prescriptions. Unfortunately, if left untreated, hypertension can have very serious long-term health consequences.

Our results provide evidence that medical drug consumption of elderly patients is highly correlated over time, and that this relationship depends on both permanent and time-varying observed and unobserved heterogeneity. Our findings also suggest that medical drug use significantly maintains future health levels and prevents transitions to worse health. In terms of policy implications, our results suggest that policies aimed at improving awareness of hypertensive diseases and the importance of the treatment of high blood pressure may help reduce cardiovascular risks, and consequent hospitalization and mortality. The presence of both state dependence and dynamics in the latent health means that also short-term policy interventions may have longer-term implications.

The remainder of this paper is organized as follows. Section 2 describes our model. Section 3 presents our approach to estimation and inference. Section 4 describes the data we use in our

empirical application. Section 5 presents our empirical results. Finally, Section 6 offers some conclusions.

2 Model

This section presents a nonlinear dynamic factor model where observable individual health outcomes are functions of observable and unobservable variables, including a latent factor, interpreted as latent health, which depends on the past values of all the variables, both observable and unobservable. One of the observable outcomes in our model is a binary indicator that switches from zero to one when a person dies. This is important because it allows us to control for endogenous sample selection due to mortality.

2.1 Elements of the model

Let $Y_{it} = (Y_{i1t}, \ldots, Y_{iJt})$ denote the vector of observable outcomes for individual $i = 1, 2, \ldots, n$ in period $t = 0, 1, \ldots, T_i$, where T_i may vary across individuals reflecting exit from the sample for a variety of reasons, in particular mortality. In our application, the first $J-1$ elements of the vector Y_{it} are observable measures of medical treatment (specifically, drug prescriptions and an indicator of hospitalization), while the last element is an indicator that switches from zero to one when a person dies. This setting is quite general because it allows the observable outcomes to be continuous or discrete. Including an indicator of mortality is especially important in order to control for endogenous sample selection due to dramatic health deterioration.

Our model consists of three basic relationships. The first reflects the assumption that each of the observable outcomes depends on a continuous latent outcome through an observation rule reflecting discretization or partial observability. Let $Y^*_{it} = (Y^*_{i1t}, \ldots, Y^*_{iJt})$ denote the vector of latent outcomes for individual i at time t. The first $J-1$ elements of Y^*_{it} may be interpreted as desired health investments, the last element as unobservable frailty. Thus, the first relationship in our model is

$$Y_{it} = h(Y^*_{it}), \tag{1}$$

where h is a known nonlinear noninvertible vector-valued function. If an outcome is fully observed (e.g., medical drug prescription), then the corresponding element of the function h is the identity function. If Y_{ijt} is instead a binary indicator (e.g., the indicator of hospitalization), then the jth element of the function h is the indicator of the event that $Y^*_{ijt} > 0$.

The second relationship is a law of motion for Y_{it}^*, relating the current value of the latent health outcomes to the current value η_{it} of a latent factor, interpreted as the latent health stock of individual i at time t,[1] the value of l time-invariant observable covariates W_i (including a constant term), the current value X_{it} of k time-varying observable covariates, past values $Y_{i,t-1}$ of the observable outcomes, and an unobservable error vector U_{it}. Assuming linearity, we specify this second relationship as

$$Y_{it}^* = A\eta_{it} + B_w W_i + B_x X_{it} + B_y Y_{i,t-1} + U_{it}, \tag{2}$$

where A is a $J \times 1$ vector of factor loadings, B_w, B_x and B_y are respectively $J \times l$, $J \times k$ and $J \times J$ matrices of unknown parameters, and the U_{it} are identically and independently distributed (iid) random vectors, distributed independently of W_i and X_{it}, with zero mean and finite nonsingular variance matrix Σ. Equation (2) formalizes the notion that the latent health outcomes are partly driven by changes in the latent health stock. It also allows the past history of medical treatment to directly affect desired health investments, reflecting habit persistence or the belief by the patient that the effectiveness of a health investment depends on continuity of the treatment. Notice that (2) allows lagged values of medical treatments to directly affect desired health investments, thus capturing complementarity or a substitutability between the various treatments. Finally, because Y_{it}^* includes frailty, our model also allows for endogenous sample selection due to mortality.

The third relationship is a law of motion for the latent factor

$$\eta_{it} = \rho\eta_{i,t-1} + C_w W_i + C_x X_{it} + C_y Y_{i,t-1} + \omega V_{it}, \tag{3}$$

where ρ is an autoregressive coefficient with $|\rho| < 1$, C_w, C_x and C_y are respectively $1 \times l$, $1 \times k$ and $1 \times J$ vectors, ω is a scale parameter, and the V_{it} are iid random health shocks, distributed independently of W_i, X_{it} and U_{it} with zero mean and unit variance. Equation (3) formalizes the notion that the time-profile of the latent health stock reflects both health deterioration and health investments (Grossman 1972, 2000; Muurinen 1982). The effect of health deterioration is captured by the autoregressive coefficient $\rho = 1 - \delta$, where δ is the rate of health depreciation, while the effects of past health investments are captured by the elements of the vector C_y associated with lagged medical treatments. In fact, lagged medical treatments can be viewed as inputs determining current health through a health production function. The role of these inputs is to restore or improve health and offset its future deterioration (preventive care). Notice that equation (3) also includes socio-demographic characteristics of an individual. Finally, because η_{it} enters the equation

[1] The model could be generalized in a straightforward manner to the case when health is treated as a multidimensional latent factor (distinguishing, for example, between physical and mental health).

for Y_{it}^*, the model allows lagged medical treatment to affect desired health investments indirectly through the latent health stock.

Equation (3) implies the following distributed lag representation of the latent factor

$$\eta_{it} = \rho^t \eta_{i0} + \frac{1-\rho^t}{1-\rho} C_w W_i + \sum_{s=0}^{t-1} \rho^s C_x X_{i,t-s} + \sum_{s=0}^{t-1} \rho^s C_y Y_{i,t-s-1} + \omega \sum_{s=0}^{t-1} \rho^s V_{i,t-s}, \tag{4}$$

where η_{i0} is health at the beginning of the period in which an individual is first observed. This representation shows that the effects of health shocks and health investments cumulate over time although, if ρ is less than 1 in modulus, events that occurred sufficiently far away in the past receive very little weight.

Given the relatively short dimension of our panel, we face an initial condition problem. We deal with it by specifying a parametric approximation for η_{i0}, the latent health stock at the beginning of the first period. Our model for η_{i0} is

$$\eta_{i0} = DZ_i + \gamma V_{i0}, \tag{5}$$

where D is a $1 \times m$ vector of unknown parameters, the vector Z_i includes both the value of time-invariant covariates and pre-sample values of Y_{it}, γ is an unknown scale parameter, and V_{i0} is a random pre-sample health shock with zero mean and unit variance. Substituting equation (4) and (5) into equation (2) gives

$$Y_{it}^* = B_{wt}^* W_i + \sum_{s=0}^{t-1} B_{xs}^* X_{i,t-s} + \sum_{s=0}^{t-1} B_{ys}^* Y_{i,t-s-1} + D_t^* Z_i + \sum_{s=0}^{t} R_s^* V_{i,t-s} + U_{it}, \tag{6}$$

where the matrices B_{wt}^*, B_{xs}^*, B_{ys}^*, D_t^* and R_s^* depend on ρ, γ, and the parameters in A, B_w, B_x, B_y, C_w, C_x, C_y and D.

Notice that the linear model (5) for the initial conditions is better viewed as an approximation, which may potentially cause specification error. Nonetheless, if $|\rho| < 1$, the influence of the initial conditions declines in later periods and becomes negligible as $t \to \infty$. Thus, there is a case for leaving a gap of S periods between the first available period and the periods used to estimate the model, that is, for estimating the model using only $Y_{i,S+1}, \ldots, Y_{i,T_i}$. The choice of S involves a trade-off between efficiency and possible misspecification bias. Actually, increasing S reduces not only the influence of initial conditions but also the amount of data used for estimation, thus alleviating the computational burden.

2.2 Relationship with other models

Because the model discussed in the previous section allows for both state dependence and dynamics in the latent factor, it can be viewed either as a generalization of dynamic panel data models or as an extension of state-space models to microeconomic panel data. As such, it encompasses as special cases several models frequently used in the literature.

Consider for simplicity the univariate version ($J = 1$) of our model. Under the restrictions $A = 0$ and $B_y = 0$, (2) becomes

$$Y_{it}^* = B_w W_i + B_x X_{it} + U_{it}.$$

This corresponds to the assumption that latent outcomes are independent over time conditional on W_i and X_{it}. We refer to this specification as the independence model.

Another special case of our general model, corresponding to the restrictions $B_y = 0$, $C_w = 0$, $C_x = 0$ and $C_y = 0$ in (2), and $\rho = 1$ and $\omega = 0$ in (3), is

$$Y_{it}^* = A\eta_i + B_w W_i + B_x X_{it} + U_{it}, \tag{7}$$

where η_i is a time-invariant individual effect. Together with (1), this is the textbook static nonlinear panel-data model with time-invariant individual effects (Baltagi 2001, Hsiao 2003). Although this specification implies that the latent outcomes are correlated over time, state dependence is spurious as it only arises from the fact that the individual effects are unobservable. Typically, there are two ways of dealing with unobservable individual effects. One is to treat them as nuisance parameters to be estimated jointly with the parameters of interest (fixed-effects approach). This approach avoids distributional assumptions but, with a fixed panel length, it suffers from an incidental parameters problem (Heckman 1981, Lancaster 2000). Another alternative consists of treating the η_i as random draws from a common distribution and integrating them out under the assumption that they are distributed independently of W_i, X_{it} and U_{it} (random-effects approach).

Relaxing the restriction that $B_y = 0$ gives yet another special case of our general model, namely

$$Y_{it}^* = A\eta_i + B_w W_i + B_x X_{it} + B_y Y_{i,t-1} + U_{it}. \tag{8}$$

This specification introduces true state dependence by allowing the latent outcome Y_{it}^* to be directly affected by the outcome observed in the previous period. When $Y_{it} = h(Y_{it}^*)$, we refer to this specification as the dynamic nonlinear panel-data model with time-invariant individual effects. Fixed-effects versions of this model have been considered by Honoré & Kyriazidou (2000), Bartolucci

& Nigro (2010) and Bartolucci, Nigro & Pigini (2018), among others. The special case when $h(\cdot)$ is the identity function corresponds to the vector autoregressive models for panel data studied by Holtz-Eakin, Newey & Rosen (1988) and Arellano & Bond (1991). Random effects versions of this model have been considered by Contoyannis, Jones & Rice (2004) among others, while an extension to the case of multivariate categorical longitudinal data has been studied by Bartolucci & Farcomeni (2009).

A second way of introducing true state dependence consists of replacing $Y_{i,t-1}$ with $Y^*_{i,t-1}$ in (8). We refer to this specification as the latent auto-regressive model with time-invariant individual effects. Pudney (2008) discusses the univariate version of this model and uses it to analyze households' perceptions of their financial wellbeing.

A third way of introducing true state dependence is to maintain the restriction $C_y = 0$ but allow the individual effects to be time-varying. One possibility is $\eta_{it} = \lambda_i^\top V_t$, where λ_i is a vector of individual-specific coefficients and V_t is a vector of time-varying factors common to all individuals. This specification, which we refer to as the interactive-effects model, includes common time trends (e.g., to capture macroeconomic effects) but allows for individual-specific reactions. Bai (2009) discuss identification and estimation of the special case when $h(\cdot)$ is the identity function under both large N and large T panel data. Another possibility is $\eta_{it} = \lambda_t^\top V_i$, where λ_t is a vector of time-varying coefficients and V_i is a vector of time-invariant individual factors. This specification is employed by Cunha et al. (2005) to separate heterogeneity from uncertainty about the returns to schooling. Eisenhauer et al. (2015) extend this model by assuming that individuals face a sequence of binary choices, allowing them to move from one schooling state to the next. Yet another possibility, corresponding to the restrictions $B_y = 0$, $C_w = 0$, $C_x = 0$, and $C_y = 0$ on (2) and (3), is the following simple autoregressive specification of the latent factor

$$\begin{aligned} Y^*_{it} &= A\eta_{it} + B_w W_i + B_x X_{it} + U_{it}, \\ Y_{it} &= h(Y^*_{it}), \\ \eta_{it} &= \rho\eta_{i,t-1} + \omega V_{it}, \end{aligned} \tag{9}$$

where V_{it} is a serially uncorrelated error distributed independently of W_i and X_{it}. We refer to this specification as the nonlinear panel-data model with time-varying individual effects. Bartolucci, Belotti & Peracchi (2015) present a simple Hausman-type test of the static model with time-invariant individual effects against this model. Notice that model (9) may be viewed as a nonlinear state-space model (Tanizaki 2003), in which η_{it} is the unobservable state variable, the first two equations correspond to the nonlinear measurement equations, and the third equation corresponds

to the state-transition equation. The model reduces to a standard linear state-space model when Y_{it}^* is fully observable.

Heiss (2008) discusses the multivariate version of model (9), whereas Heiss (2011) presents an application to jointly modeling self-rated health and mortality. In his application, Y_{it}^* consists of observable self-rated health and unobservable frailty, η_{it} is the latent health stock, and death occurs if frailty exceeds a threshold. Notice that his model implies that

$$\eta_{it} = \rho^t \eta_{i0} + \omega \sum_{s=0}^{t-1} \rho^s \omega V_{i,t-s},$$

where η_{i0} is initial health. Thus, an undesirable model feature in the context considered by Heiss (2011) is that the health stock evolves outside of an individual's control as a weighted-sum of past health shocks, with geometrically declining weights. This is in sharp contrast with standard models in health economics, such as the model of Grossman (1972), where current health is a function of both past health and current health investments, as in equation (3).

2.3 Identification

The model consisting of (1), (2), (3) and (5) is heavily parameterized, so identification is an important issue. The conditions for parametric identification also play an essential role in the model estimation strategy discussed in Section 3.

As usual in factor analysis, we assume that the latent outcomes are mutually independent conditional on the observable covariates X_{it}, the common latent factor η_{it}, and the pre-sample individual shock V_{i0}. Hence, the variance matrix Σ is diagonal, with diagonal elements $\sigma_j^2 = \text{Var}(U_{ijt})$, $j = 1, \ldots, J$. To obtain a tractable parametric model, we assume U_{it} and V_{it} to be iid Gaussian.

Moreover, as usual in latent variables models, our model requires that each latent variable is assigned a location and a scale for identification. This is true for the latent outcomes in the vector Y_{it}^*, as well as for the latent health stock η_{it}. For the binary outcomes we use the convention of setting the threshold to zero and the errors variances to one. As a standard practice in factor analysis, location and scale of η_{it} is normalized by setting the intercept in (3) to zero and the variance ω^2 of the random health shocks to one. With these normalizations the variance of the random term in the initial conditions equation is equal to γ^2. Thus, V_{i0} is not normalized to have unit variance. The total number of free parameters in our model is therefore equal to $p = (J+1)(1+l+k+J)+m$. We refer to these parameters as the structural parameters and represent them as elements of a vector

ψ in a parameter space Ψ, a subset of the p-dimensional real Euclidean space. From the reduced form in (6) we have a total of $q = J + JT(m + l + k + J + 1)$ estimable parameters. We refer to these parameters as the reduced-form parameters and represent them as elements of a vector θ in a parameter space Θ, a subset of the q-dimensional real Euclidean space. Note that while both p and q depend on the number J of observable outcomes and on m, l and k, the number T of time periods affects only q. Further, identification of ρ requires at least three panel waves.

Also note that if time-invariant variables W_i are included in the equation (5) for initial health, then their coefficients in D are not separately identified from the coefficients C_w in the equation (3) for current health. Thus, in this case we would set the coefficients C_w in equation (3) to zero. Moreover, since time-varying variables X_{it} and $Y_{i,t-1}$ are included both in the observable outcomes equations (2) and in the equation (3) for current health, we can interpret their parameters B_x and B_y in the former equations as their direct effects on the outcomes, and their parameters C_x and C_y in the latter equation as their indirect effects – through the latent factor. Separate identification of these direct and indirect effects crucially relies on the fact that both effects enter parameters B^*_{x0} and B^*_{y0} in the reduced form (6), while parameters B^*_{xs} and B^*_{ys}, with $s = 1, \ldots, t-1$, only depends on the indirect effects.

Finally, note that, as mentioned above, the first element of Y_{it} may consist of a selection indicator which is equal to 1 if a unit drops out of the sample from time $t+1$ and 0 otherwise. In this case, the first element of C_y and D_y, and the first column of B_y and B^*_{ys} are necessarily set to zero.

3 Estimation

We propose to estimate the model in Section 2 by combining the maximum simulated likelihood (MSL) and the minimum distance (MD) methods. In what follows we denote by $\theta_0 \in \Theta$ the true value of the vector of q reduced-form parameters in equation (6) and by $\psi_0 \in \Psi$ the true value of the vector of p structural parameters in equations (2), (3) and (5).

3.1 Estimation of the reduced-form parameters

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i,T_i})$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{i,T_i})$, and $\mathbf{V}_i = (V_{i0}, V_{i1}, \ldots, V_{i,T_i})$ respectively denote the histories of the observable outcomes, the observable time-varying covariates, and the latent health shocks for individual i. Under our assumptions, conditional on W_i, $\mathbf{X}_i$, Z_i, $\mathbf{V}_i$, the observable outcomes $\mathbf{Y}_i$ are independent over time. Hence, the conditional density of $\mathbf{Y}_i$ given $W_i = w_i$,

$\mathbf{X}_i = \mathbf{x}_i$, $Z_i = z_i$ and $\mathbf{V}_i = \mathbf{v}_i$, with $\mathbf{x}_i = (x_{i1} \ldots, x_{i,T_i})$ and $\mathbf{v}_i = (v_{i0}, \ldots, v_{i,T_i})$, is of the form

$$f(\mathbf{y}_i \,|\, w_i, \mathbf{x}_i, z_i, \mathbf{v}_i; \theta) = \prod_{t=1}^{T_i} f(y_{it} \,|\, y_{i,t-1}, w_i, x_{it}, z_i, v_{it}; \theta),$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,T_i})$, $f(y_{it} \,|\, y_{i,t-1}, w_i, x_{it}, z_i, v_{it}; \theta)$ denotes the conditional density of Y_{it} given $Y_{i,t-1} = y_{i,t-1}$, $W_i = w_i$, $X_{it} = x_{it}$, $Z_i = z_i$ and $V_{it} = v_{it}$, and θ is the vector of parameters entering the model consisting of (2), (3) and (5).

Under the assumption that the observable outcomes are conditionally independent given $Y_{i,t-1}$, W_i, X_{it}, Z_i, and V_{it}, the conditional density of Y_{it} is simply the product of the conditional densities of the J outcomes

$$f(y_{it} \,|\, y_{i,t-1}, w_i, x_{it}, z_i, v_{it}; \theta) = \prod_{j=1}^{J} f_j(y_{ijt} \,|\, y_{i,t-1}, w_i, x_{it}, z_i, v_{it}; \theta), \tag{10}$$

where $f_j(y_{ijt} \,|\, y_{i,t-1}, w_i, x_{it}, z_i, v_{it}; \theta)$ denotes the conditional density of the jth outcome Y_{ijt} given $Y_{i,t-1} = y_{i,t-1}$, $W_i = w_i$, $X_{it} = x_{it}$, $Z_i = z_i$ and $V_{it} = v_{it}$. Since the health shocks $\mathbf{V}_i$ are unobservable, they must be integrated out of $f(\mathbf{y}_i \,|\, w_i, \mathbf{x}_i, z_i, \mathbf{v}_i; \theta)$. The resulting conditional density of $\mathbf{Y}_i$ given $W_i = w_i$, $\mathbf{X}_i = \mathbf{x}_i$ and $Z_i = z_i$ is

$$f(\mathbf{y}_i \,|\, w_i, \mathbf{x}_i, z_i; \theta) = \int f(\mathbf{y}_i \,|\, w_i, \mathbf{x}_i, z_i, \mathbf{v}; \theta)\, g(\mathbf{v})\, d\mathbf{v},$$

where $g(\mathbf{v})$ is the joint density of the T_i+1 latent health shocks, which are assumed to be distributed independently of W_i, $\mathbf{X}_i$ and Z_i. The implied log-likelihood function for a random sample of n individuals, each observed for T_i periods, is

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n} \ln f(\mathbf{Y}_i \,|\, W_i, \mathbf{X}_i, Z_i; \theta) \\ &= \sum_{i=1}^{n} \ln \left[\int f(\mathbf{Y}_i \,|\, W_i, \mathbf{X}_i, Z_i, \mathbf{v}; \theta)\, g(\mathbf{v})\, d\mathbf{v} \right]. \end{aligned} \tag{11}$$

A maximum likelihood (ML) estimator of the population parameter θ_0 is a maximizer $\hat{\theta}_n$ of $L_n(\theta)$ over Θ. As $n \to \infty$, $\hat{\theta}_n$ is consistent for θ_0 and asymptotically Gaussian under mild regularity conditions (see, e.g, Gouriéroux & Monfort 1997). The multivariate integral in (11) can in principle be evaluated by multivariate numerical integration methods, such as Gaussian quadrature on a sparse grid (Heiss & Winschel 2008). Another alternative is nonlinear Kalman filters (Durbin & Koopman 2001), for which Heiss (2008) proposes a convenient sequential Gaussian quadrature method.

The MSL method is simpler because it does not require evaluating the multivariate integral in (11). Instead, it maximizes over Θ the following approximation to $L_n(\theta)$, obtained replacing the multivariate integral by a sample average,

$$L_n^*(\theta) = \sum_{i=1}^{n} \ln \left[\frac{1}{R} \sum_{r=1}^{R} f(\mathbf{Y}_i \,|\, W_i, \mathbf{X}_i, Z_i, \mathbf{V}_i^r; \theta) \right],$$

where $\mathbf{V}_i^1, \ldots, \mathbf{V}_i^R$ are R draws from the multivariate distribution $g(\mathbf{v})$ of the health shocks. Under mild regularity conditions, $L_n^*(\theta)$ convergence in probability, uniformly in θ as $R \to \infty$, to the log-likelihood function (11). As a result, if the number of draws increases with n and $n \to \infty$, a maximizer $\tilde{\theta}_n$ of $L_n^*(\theta)$ is consistent for θ_0, asymptotically Gaussian, and asymptotically equivalent to a ML estimator $\hat{\theta}_n$ (Hajivassiliou & Ruud 1994; Gouriéroux & Monfort 1997).

Given the assumption on the V_{it}, we could generate the elements of the vector $\mathbf{V}_i^r$ as pseudo-random draws from a multivariate standard Gaussian distribution. Instead, we base our simulations on Halton sequences (Train 2003), as they provide a more systematic coverage of the domain of integration than pseudo-random draws. This usually results in less draws, smaller integration error, and faster convergence rates. Since Halton sequences are deterministic, following Wang & Hickernell (2000) we introduce randomness by employing a random start procedure. Specifically, we first randomly draw an integer N_0 between 0 and some large value K, and then create a Halton sequence starting from N_0. The simulated log-likelihood function of factor models like ours may have multiple local maxima. For this reason, we use multiple starting values and check whether we end up with the same parameter estimates.

3.2 Estimation of the structural parameters

The MD approach to estimation of the structural parameters of our model relies on the fact that the population parameters θ_0 and ψ_0 are linked through the nonlinear relationship $\theta_0 = g(\psi_0)$, where $g\colon \Psi \to \Omega$ is a differentiable function with Jacobian matrix G. For (local) identifiability, we also need $G(\psi)$ to be of full rank in an open neighborhood of ψ_0. Appendix A presents the structure of g and G, and illustrates with a simple special case.

Given a MSL estimate $\tilde{\theta}_n$, the MD method suggests estimating the vector ψ_0 of structural parameters by picking an element of Ψ such that the vector $\tilde{\theta}_n - g(\psi)$ is smallest in a suitably chosen metric. The resulting estimator of ψ_0 is unique, $\sqrt{n}$-consistent and asymptotically Gaussian under general conditions (Ferguson 1996). An asymptotically optimal MD estimator is a solution

$\hat{\psi}_n$ to the problem

$$\min_{\psi \in \Psi} Q_n(\psi) = [\tilde{\theta}_n - g(\psi)]^\top \tilde{V}_n^{-1} [\tilde{\theta}_n - g(\psi)], \tag{12}$$

where the $q \times q$ matrix $\tilde{V}_n$ is a positive definite estimate of the asymptotic variance of $\tilde{\theta}_n$. Under general conditions, $\sqrt{n}(\hat{\psi}_n - \psi_0)$ converges in distribution, as $n \to \infty$, to a multivariate normal distribution with mean zero and variance matrix $(G_0 V_0^{-1} G_0^\top)^{-1}$, where $G_0 = G(\psi_0)$ denotes the $p \times q$ Jacobian matrix of g evaluated at ψ_0 and V_0 denotes the asymptotic variance of $\hat{\theta}_n$.

The estimator $\hat{\psi}_n$ can be computed using an iterative procedure. Specifically, starting from an initial estimate $\hat{\psi}^{(0)}$, the updated estimate at the $(h+1)$th iteration is given by

$$\hat{\psi}^{(h+1)} = (\hat{G}_h \hat{V}_n^{-1} \hat{G}_h^\top)^{-1} \hat{G}_h \hat{V}_n^{-1} (\hat{\theta}_n - \hat{g}_h + \hat{G}_h^\top \hat{\psi}^{(h)}), \qquad h = 0, 1, \ldots,$$

where $\hat{G}_h = G(\hat{\psi}^{(h)})$ and $\hat{g}_h = g(\hat{\psi}^{(h)})$. This corresponds to a GLS regression of the transformed reduced-form estimates $\hat{\theta}_n - \hat{g}_h + \hat{G}_h^\top \hat{\psi}^{(h)}$ on the columns of $\hat{G}_h$ with weighting matrix $\hat{V}_n^{-1}$.

4 Data

We estimate the model presented in Section 2 using the longitudinal data set employed by Atella et al. (2006). This data set – constructed by combining a number of administrative archives maintained by one regional unit of the Italian national health-care system – contains very detailed micro-level information on medical drug use, hospitalization, and mortality. Because our data contain objective measures of health from administrative health-care records, they offer an important advantage relative to studies based on self-reported measures of health. For the reasons discussed in the Introduction, we focus on hypertensive patients treated with active ingredients in five therapeutic main groups (TMG) of the Anatomical Therapeutical and Clinical (ATC) Classification System, namely antihypertensives, diuretics, beta blocking agents, calcium channel blockers, and agents acting on the renin-angiotensin system. Our observable outcomes consist of medical drug use, treated as a continuous variable, and binary indicators for hospitalization by diagnosis-related group (DRG) and death. The inclusion of the latter is especially important, not only because it is an objective – albeit extreme – measure of health, but also because it allows us to control for endogenous sample selection due to mortality.

4.1 Construction of the data set

Our data set has been constructed by combining three administrative registries maintained by the Pharmaceutical Service Department of the local health authority covering the southern part

of the Italian province of Treviso, one of the prosperous provinces of North-Eastern Italy. This allows us to obtain a very rich, and to our knowledge unique, longitudinal data set containing very detailed information on a representative sample of hypertensive patients treated with active ingredients in five therapeutic main groups (TMG) of the Anatomical Therapeutical and Clinical (ATC) Classification System, namely antihypertensives, diuretics, beta blocking agents, calcium channel blockers, and agents acting on the renin-angiotensin system.

Specifically, the first administrative registry is the drug prescription database, which contains records of patient prescriptions, including the date the prescription is made, the ATC code of the substance prescribed, the number of packages prescribed, and the unit price of the package. The unit price of the package allows us to recover information about the number of pills contained in each package. This registry also includes gender and date of birth of the patient receiving the medications, a unique anonymized patient identifier, a unique anonymized identifier of the physician who prescribes the medication, and gender and date of birth of the physician.

The second administrative registry is the hospitalization registry, which contains records of each single hospitalization episode, including date of entry and dismissal, primary DRG, and cost of hospitalization. Through the anonymized personal identifiers, we link patient prescription and hospitalization information to the last registry, the death and transfer registry. The resulting dataset allows us to follow individual patients through all their accesses to public health-care services until they either die or leave the local health authority. Complete data are available for all registries from 1997 to 2002.

Relative to survey data, these administrative data have both advantages and disadvantages. An important advantage is that they do not present problems which are typical of survey data, namely unit and item non-response, measurement errors, and bias effects due to interaction with interviewers. Another advantage is that they contain extremely rich information on health-care services received by patients. Further, mobility is extremely low in our data and sample selection is (almost) entirely due to mortality. The main disadvantage is that they contain little information on a patient's socio-economic characteristics. In particular, no information on income and education is available.

4.2 Outcomes and controls

We consider three observable outcomes (thus, $J = 3$): medical drug, hospitalization, and mortality. While drug use is a continuous variables, the other two are binary indicators. Since people who

die drop out of the sample, the equation for the first outcome can be interpreted as an endogenous selection equation. The time unit is the year.

We observe whether an individual is alive at the end of period t, and whether she has been hospitalized during period t. The associated outcome variables are binary 0-1 indicators. Finally, medical drug use C_{it} is computed by average daily drug purchase,[2] computed as the total number of pills bought by patient i during period t divided by the number of days T_{it} the patient is observed during the period. Thus

$$C_{it} = \frac{1}{T_{it}} \sum_{j=1}^{J_{it}} N_{ij} \cdot P_{ij},$$

where J_{it} is the number of prescriptions filled to patient i during period t, and P_{ij} and N_{ij} are respectively the number of pills per pack and the number of packs contained in the jth prescription. Since patients do not buy medical drugs when they are hospitalized or after they are dead, T_{it} is given by the total number of days the patient is alive minus the number of days a patient is hospitalized during such period.

Our set of control variables is quite limited and includes age and gender, an indicator for being exempt from the payment of the prescription charge, an indicator for the gender of the patient, and one for concordance between the gender of the patient and the gender of the physician prescribing the drugs (physician and patient gender concordance). Specifically, the indicator for being exempt is equal to one if an individual is exempt from the co-payment prior to 2001 (the year when co-payment was temporarily abolished in Italy). Note that all patients in our sample were either always or never exempt from the co-payment in all the sample years prior to 2001. Thus, this indicator is time-invariant for all individuals. Patients in our sample were exempt from the ticket either because of low income, or because they had disability or specific chronic rare pathologies diagnosed by specialists. Thus, although the rules for exemption are somewhat complicated, including this indicator allows us to somehow control for patients' income. Finally, note that the gender of the physician is constant over time for about 98.1% of the patients observed in our sample. To reduce the computational burden, we treated this indicator as time-invariant for all individuals by assigning the gender of the physician who made most of the prescriptions for the few patients who faced a change.

We consider all patients who are present in the administrative archives in 1997, were born

[2] Patients need not consume all the medical drug they purchase. Further, they may buy medical drug in a pharmacy outside the boundaries of their local health authority. Because these events appear to be rare, we do not worry about this source of measurement error.

between 1920 and 1939 (aged 61–80 in year 2000),[3] and received at least one prescription during the period of a year.[4] Since some of the medications considered here are also employed against kidney failure, we drop patients who were hospitalized for renal diseases but never for cardiovascular diseases (2.7% patients dropped), patients with missing data on any of the variables used (3.4% patients dropped), and patients with an average drug use greater than 5 pills per day (2.9% patients dropped). The last selection criterion reflects the fact that these outliers may represent cases of co-morbidity. Finally, to apply our dynamic model we keep only patients observed for at least four consecutive periods (2.5% dropped) The final sample consists of 4,296 patients observed for 5.9 years on average. Figure 1 shows the distribution of the patients included in our final sample by gender and year of birth.

4.3 Descriptive statistics and preliminary evidence

Because we expect a possibly quite different relation between health and medical care depending on gender, we carry out our analysis separately for males and females. Table 1 shows descriptive statistics for our sample. Mortality, hospitalization and medical drug use are all higher for men than for women. Most of the physicians in our sample are men. Finally, the percentage of individuals who are exempt from paying the prescription charge is very low.

Figure 2 shows nonparametric kernel estimates of the density of medical drug use by gender. Both densities display two peaks corresponding respectively to about half a pill per day and 1 pill per day. Figure 3 shows the age-profile of average drug use, hospitalization and mortality by gender. All measures increase with age and are higher for men at all ages.

5 Results

Our observable health outcomes are mortality, hospitalization and medical drug use (hence, $J = 3$). The corresponding vector of latent outcomes is $Y_{it}^* = (Y_{i1t}^*, Y_{i2t}^*, Y_{i3t}^*)'$, where Y_{i1t}^* is the logarithm of medical drug use, Y_{i2t}^* is the propensity to be hospitalized, and Y_{i3t}^* is latent "frailty". We think of all these latent outcomes as driven by the observable covariates and the latent health stock. Although mortality is an extreme measure of health, it is an objective measure. As such, it is not subject to problems like reporting bias or subjective response scales, that typically affect other subjective measures like self-reported health or conditions. Moreover, the inclusion of this

[3] We restrict attention to these cohorts because they comprise the bulk of the population suffering of hypertension.

[4] Patients who do not receive prescriptions for more than 365 days are likely to be affected by mild hypertension that could be treated simply by a healthy diet and by reducing stress factors.

outcome is especially important because it allows us to control for endogenous sample selection due to mortality. Given the peculiarities of our sample, panel attrition is (almost) entirely due to mortality.[5]

Except for Y^*_{i1t}, which is always observable, we only observe binary indicators corresponding to Y^*_{i2t}, and Y^*_{i3t}. Thus, the elements of the vector Y_{it} of observable outcomes are

$$Y_{ijt} = \begin{cases} Y^*_{ijt}, & \text{if } j = 1, \\ 1[Y^*_{ijt} > 0], & \text{if } j = 2, 3, \end{cases} \tag{13}$$

where $1[A]$ is the indicator function of the event A.

5.1 Parameter estimates

We estimate the reduced-form parameters in (6) by MSL. The maximization routine that computes the MSL estimates for these models is written in Mata, the matrix programming language of the statistical package Stata, and is based on the Newton-Raphson algorithm with numerical first and second derivatives. The MSL estimates are based on 50 draws, using Halton sequences, from the multivariate standard Gaussian distribution of the unobservable health shocks. To reduce the influence of the initial conditions and the computational burden, we leave a gap of $S = 2$ periods after the initial period 0, so we estimate the model with $t = 3, \ldots, T_i$ periods.[6]

The time-varying variables included in our model are age, an indicator for being hospitalized in $t - 1$, and the logarithm of medical drug use in $t - 1$. The time-invariant variables included in our model are an indicator for gender concordance between the patient and the physician, and an indicator for being exempt from the co-payment prior to 2001. The variables in the initial conditions, which may be viewed as additional instruments, include the logarithm of medical drug use in the first sample period, an indicator for being hospitalized in the first sample period, an indicator for gender concordance between the patient and the physician, and an indicator for being exempt from the co-payment prior to 2001. Appendix B presents the estimated coefficients of the reduced-form parameters by gender.

Table 2 shows MD estimates of the structural parameters by gender. The iterative routine that computes the MD estimates for these models is also written in Mata, the matrix programming language of Stata. Estimated factor loadings of all heath outcomes have a positive sign, denoting that outcomes are all positively related to the latent common factor. Because these outcomes,

[5] The only other reason for losing a unit sample is mobility, but this is extremely rare in our data.

[6] Estimation of model (6) required about 5 days for the subsample of men and 9 days for the subsample of women on an Intel Dual Core i5-6500 CPU @ 3.20GHz computer running Windows 10 and Stata Version 15.

namely mortality, hospitalization, and medical drug use, are all expected to increase as health decreases, we interpret the latent common factor η as a measure of bad health. The importance of modeling the latent health stock jointly with the health outcomes is to capture correlation in the error terms associated with endogenous medical care inputs that affect health. Estimation of these factor loadings are an important feature of our model, allowing for joint estimation of endogenous medical care use and health outcomes.

The top panel of Table 2 shows the estimated coefficients of the equation (5) for the initial conditions. Not surprisingly, the initial stock of bad health is positively related to both hospitalization and the level of medical drug use in the initial period, though the latter is not statistically significant for women. Being exempt from co-payment is positive and significant for women. Note that the exemption indicator is a proxy for either the presence of disability or low income. Having a physician of the same gender is positively related to the initial stock of bad health for men, while is negatively related for women. Thus a female physician prescribing medical drugs is associated to better initial health for both male and female patients. The scale parameter γ in equation (5) is highly statistically significant for women but not for men.

The second panel of Table 2 shows the estimated coefficients of the latent health equation, while the following three panels show the estimated coefficients of the three equations for the observable outcomes. Note again that including the mortality equation is another important feature of our model, allowing for endogenous sample selection. Recall that when a variable is included in both the equation for the latent health stock and the equations for the observed health outcomes, we can interpret its coefficient in the latter equations as the direct effect on the outcomes, and its coefficient in the former as the indirect effect – through the latent health stock. The autoregressive coefficient ρ captures health deterioration. This coefficient is equal to .367 for men and to .621 for women, suggesting a higher rate of health depreciation $(1 - \rho)$ for men.

In addition to health deterioration, captured by ρ, an age-related decline in the latent health stock is also reported for men. This means that age indirectly increases the probability of death and hospitalization, through latent health. The estimated direct effects of age are not statistically significant for men. For women, a reduction of bad health as age increases is reported to mitigate health depreciation captured by ρ. This means that age indirectly decreases the observable health outcomes for women. Nonetheless, these negative indirect effects are dominated by the significant and positive direct effects of age on the probability of death and hospitalization and, to a lesser extent, on medical drug use.

Being exempt from payment of the prescription charge increases both the probability of death and hospitalization. This effect is what one would expect, because this indicator is a proxy for either the presence of disability or low income. The partial effect on medical drug use is not significant. A male GP is associated with a lower probability of hospitalization both for male and female patients.

Being hospitalized in period $t-1$ significantly reduces health, especially for men. In fact, the estimated coefficients of $Y_{2,t-1}$ on the latent bad health equation is positive. For men, lagged hospitalization has also a negative direct effects on the current probability of death and on the probability of being hospitalized again in the current period. One interpretation is that an hospitalization episode makes it less likely to be hospitalized again or even dying in the following period, possibly leading to a temporary health restore, but also reduces health, with negative indirect effects in the future years. This is consistent with findings from Yang, Gilleskie & Norton (2009). For women, the direct effect of previous period hospitalization on the current probability of death is also negative but not significant. For women, lagged hospitalization is instead positively correlated with the probability of being hospitalized again in the current period. Thus, for women, both the direct and the indirect effects of a hospitalization episode go in the same direction of increasing the future probability of a subsequent hospitalization. Finally, hospitalization in a previous year suggests a slightly lower use of medical drugs in the the following year, although the coefficient is not significant for men.

Next, we investigate the effects of lagged medical drug use on health production and on current observable outcomes and medical use. First of all, lagged medical drug use significantly reduces latent bad health, especially for men. Thus, as expected, medical drug consumption acts as a health investment. In fact, as long as differences in purchases reflect differences in consumption levels, additional prescription drug use may prevent transitions to worse health.

Not surprisingly, lagged medical drug use significantly affects medical consumption today. Specifically, previous prescription drug is positively correlated with contemporaneous drug consumption. This may reflect both the fact that the specific illness considered here, i.e. hypertension, is a chronic pathology, but also possibly the habitual or dependent nature of medical care use at older ages. Moreover, individuals who used more prescription drugs in the previous year are more likely to be hospitalized this year. These estimates suggest that previous medical drug use has a direct effect on current use and hospitalization, independent of its indirect effect through changes in health. Lagged medical drug use has also a positive direct effect increasing the current probability of death for men. Nonetheless, this direct effect is then contrasted by an indirect effects through

health going in the opposite direction. For women, both direct and indirect effects of lagged medical drug use go in the same direction, reducing the future probability of death.

5.2 Direct, indirect and total partial effects

The estimated coefficients in the equations for mortality and hospitalization reported in Table 2 are not very easy to interpret. Table 3 reports estimated effects from the dynamic factor model by gender. Specifically, we report estimates of the direct, indirect and total marginal effects implied by our model. These effects are reported for lagged hospitalization, lagged medical drug use, and age. Standard errors are based on the delta method. For the probabilities of hospitalization (Y_2) and death (Y_3), these are the partial effects for the baseline individual. The baseline individual is a man or woman aged 75 years, not exempt from the prescription charge, with a GP not of the same gender, who in $t-1$ had an average medical consumption of one pill per day and was not hospitalized. For discrete variables, such lagged hospitalization, these are obtained by taking differences. For continuous variables, such as lagged medical drug use and age, they are obtained by taking the derivative of probabilities with respect to the variable in question. For the sake of completeness we also report direct, indirect and total marginal effects for medical drug use (Y_1).

Not surprisingly, both the probability of death and hospitalization increase with age. Looking at the total effects, all else equal, individuals are 0.1% more likely to die as they get one year older. Each additional year, the probability of hospitalization increases 0.6% for men and 0.4% for women. Being hospitalized in $t-1$ increases the probability of death by 1% for men and by 0.7% for women. This is mainly due to the "long-run" indirect effects of hospitalization reducing latent health. Lagged hospitalization strongly increases the probability of being hospitalized again in the current period, despite the smaller direct effect reducing such a probability for men. The indirect effect of lagged medical use preserving or restoring latent health is slightly overwhelmed by the direct effect in the opposite direction both on the probability of death and hospitalization. The only exception is the probability of death for women, where direct and indirect effects are both negative. In this case, an increase in medical use reduces such a probability by 0.2% in the next period. Finally, for medical drug use, estimated effects actually show the strong persistence of medical consumption, typical of a chronic condition such as hypertension, under consideration here. The total partial effects of lagged hospitalization and age are not significant.

6 Conclusions

In this paper, we present a tractable dynamic factor model of health and medical treatment, where observable health outcomes are driven by the latent individual health and lagged medical treatment. This paper extends the existing literature on the dynamic relationship between health and medical treatment in several ways. The dynamics of health arises from both exogenous health depreciation and endogenous health investments. Our model allows us investigating the effect of medical treatment on health production, and on future medical use and observable outcomes. By including mortality among observable outcomes, our model accounts for endogenous attrition. Because we allow for both state dependence and dynamics in the latent factor, our model can be viewed as a generalization of dynamic panel data models or as an extension of state-space models to microeconomic panel data. Thus, because of its generality, our method can be also applied in a variety of other settings.

We estimate the model by maximum simulated likelihood and minimum distance methods using a rich longitudinal data set containing very detailed information on medical drug use, hospitalization, and mortality for a representative sample of hypertensive individuals. In particular, we focus on patients who were prescribed medical drugs employed against hypertension, a chronic asymptomatic pathology affecting a large share of the adult population. Due to the peculiarity of the pathology considered here, focusing on such patients offers the possibility of having helpful insights into the relation between health and medical care.

Our model allows us investigating the effect of medical drug use on health production, and on future medical use and observable outcomes. Our study provides evidence that medical care consumption is highly correlated over time, and that this relationship depends on both permanent and time-varying observable and unobservable heterogeneity. Moreover, our study produces estimates of both direct and indirect effects of medical care use on observable health outcomes. These indirect effects can be viewed as longer-term effects through latent health. First of all, our findings suggest that, medical drug use significantly maintains future health levels and prevents transitions to worse health. Many previous analyses using simple empirical models to analyse the relationship between medical care and health were not successful in this direction, because of simultaneity through the unobservable components of health deterioration, failing to recognize the health investment role of medical care. Considering observable health outcomes, these indirect effects of medical care use have a clear role in reducing the future probability of hospitalization and death.

In terms of policy implications, our results suggest that policies aimed at improving awareness

of hypertensive diseases and the importance of the treatment of high blood pressure may help reduce cardiovascular risks, and consequent hospitalization and mortality. The presence of both state dependence and dynamics in the latent health means that also short-term policy interventions may have longer-term implications.

References

Arellano, M. & Bond, S. (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies* **58**(2), 277–297.

Atella, V., Depalo, D., Peracchi, F. & Rossetti, C. (2006), 'Drug compliance, co-payment and health outcomes: Evidence from a panel of Italian patients', *Health Economics* **15**(9), 875–892.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Bai, J. & Li, K. (2016), 'Maximum likelihood estimation and inference for approximate factor models of high dimension', *Review of Economics and Statistics* **98**(2), 298–309.

Bai, J. & Wang, P. (2014), 'Identification theory for high dimensional static and dynamic factor models', *Journal of Econometrics* **178**(2), 794–804.

Baltagi, B. (2001), *Econometric Analysis of Panel Data (2nd ed.)*, John Wiley & Sons, New York.

Bartolucci, F., Belotti, F. & Peracchi, F. (2015), 'Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data', *Journal of Econometrics* **184**(1), 111–123.

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.

Bartolucci, F. & Nigro, V. (2010), 'A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator', *Econometrica* **78**(2), 719–733.

Bartolucci, F., Nigro, V. & Pigini, C. (2018), 'Testing for state dependence in binary panel data with individual covariates by a modified quadratic exponential model', *Econometric Reviews* **37**(1), 61–88.

Case, A. & Deaton, A. (2005), Broken down by work and sex: How our health declines, *in* D. A. Wise, ed., 'Analyses in the Economics of Ageing', University of Chicago Press, Chicago, pp. 185–212.

Contoyannis, P., Jones, A. M. & Rice, N. (2004), 'The dynamics of health in the British Household Panel Survey', *Journal of Applied Econometrics* **19**(4), 473–503.

Cunha, F., Heckman, J. & Navarro, S. (2005), 'Separating uncertainty from heterogeneity in life cycle earnings', *Oxford Economic Papers* **57**(2), 191–261.

Durbin, J. & Koopman, S. J. (2001), *Time Series Analysis by State Space Methods*, Vol. 38, Oxford University Press, New York.

Eisenhauer, P., Heckman, J. J. & Mosso, S. (2015), 'Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments', *International Economic Review* **56**(2), 331–357.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, London.

French, D. & O'Hare, C. (2013), 'A dynamic factor approach to mortality modeling', *Journal of Forecasting* **32**(7), 587–599.

Geweke, J. (1977), The dynamic factor analysis of economic time series models, *in* D. Aigner & A. Goldberger, eds, 'Latent variables in socio-economic models', North-Holland, pp. 365–383,.

Gouriéroux, C. & Monfort, A. (1997), *Simulation-based Econometric Methods*, Oxford University Press, New York.

Grossman, M. (1972), 'On the concept of health capital and the demand for health', *Journal of Political Economy* **80**(2), 223–255.

Grossman, M. (2000), The human capital model, *in* A. J. Culyer & J. P. Newhouse, eds, 'Handbook of Health Economics', Vol. 1, Elsevier, Amsterdam, pp. 347–408.

Hajivassiliou, V. A. & Ruud, P. A. (1994), Classical estimation methods for LDV models using simulation, *in* R. Engle & D. McFadden, eds, 'Handbook of Econometrics', Vol. 4, Elsevier, New York, pp. 2383–2441.

Heckman, J. J. (1981), The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, *in* C. F. Manski & D. L. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT Press, Cambridge, MA.

Heiss, F. (2008), 'Sequential numerical integration in nonlinear state space models for microeconometric panel data', *Journal of Applied Econometrics* **23**(3), 373–389.

Heiss, F. (2011), 'Dynamics of self-rated health and selective mortality', *Empirical Economics* **40**(1), 119–140.

Heiss, F. & Winschel, V. (2008), 'Likelihood approximation by numerical integration on sparse grids', *Journal of Econometrics* **144**(1), 62–80.

Holtz-Eakin, D., Newey, W. & Rosen, H. S. (1988), 'Estimating vector autoregressions with panel data', *Econometrica* **56**, 1371–1395.

Honoré, B. E. & Kyriazidou, E. (2000), 'Panel data discrete choice models with lagged dependent variables', *Econometrica* **68**(4), 839–874.

Hsiao, C. (2003), *Analysis of Panel Data (2nd ed.)*, Cambridge University Press, New York.

Lancaster, T. (2000), 'The incidental parameter problem since 1948', *Journal of Econometrics* **95**(2), 391–413.

Muurinen, J.-M. (1982), 'Demand for health: A generalised Grossman model', *Journal of Health Economics* **1**(1), 5–28.

Pudney, S. (2008), 'The dynamics of perception: Modelling subjective wellbeing in a short panel', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1), 21–40.

Sargent, T. J. & Sims, C. A. (1977), 'Business cycle modeling without pretending to have too much a priori economic theory', *New methods in Business Cycle Research* **1**, 145–168.

Tanizaki, H. (2003), Nonlinear and non-Gaussian state-space modeling with Monte Carlo techniques: A survey and comparative study, *in* D. Shanbhag & C. Rao, eds, 'Handbook of Statistics', Vol. 21, Elsevier, New York, pp. 871–929.

Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, UK.

Wang, X. & Hickernell, F. J. (2000), 'Randomized Halton sequences', *Mathematical and Computer Modelling* **32**(7-8), 887–899.

Yang, Z., Gilleskie, D. B. & Norton, E. C. (2009), 'Health insurance, medical care, and health outcomes: A model of elderly health dynamics', *The Journal of Human Resources* **44**(1), 47–114.

Table 1: Descriptive statistics

	Men		Women	
	Mean	SD	Mean	SD
Log. of medical drug use	0.097	0.643	0.012	0.647
Mortality	0.013	0.115	0.008	0.088
Hospitalization	0.214	0.410	0.179	0.383
Age	69.79	5.811	70.76	5.825
Exempt from charge	0.025	0.157	0.013	0.115
Same gender physician	0.849	0.358	0.159	0.366
Observations	10,143		15,338	
Patients	1,717		2,579	

Table 2: Estimated coefficients of the structural parameters by gender (* significant at 10%; ** significant at 5%; *** significant at 1%)

	Men	Women
Initial conditions (bad health) η_0		
Y_{10}	0.452***	0.035
Y_{20}	2.390***	1.314***
Gender concordance	0.333***	-0.495***
Exempt	-0.275	0.974***
γ	0.002	1.598***
Latent common factor (bad health) η		
Age - 75	0.057***	-0.018***
$Y_{1,t-1}$	-0.348***	-0.067***
$Y_{2,t-1}$	1.737***	0.624***
ρ	0.367***	0.621***
Log medical drug use Y_1		
Constant	0.033	0.059***
Gender concordance	0.033	0.024
Exempt	0.064	-0.081
Age - 75	-0.002	0.003*
$Y_{1,t-1}$	0.688***	0.720***
$Y_{2,t-1}$	-0.052	-0.036*
Factor loading	0.014	0.026*
Hospitalization Y_2		
Constant	-0.890***	-1.600***
Gender concordance	-0.071***	0.229***
Exempt	0.484***	0.132
Age - 75	-0.007	0.054***
$Y_{1,t-1}$	0.243***	0.182***
$Y_{2,t-1}$	-0.330***	0.136***
Factor loading	0.506***	0.983***
Death Y_3		
Constant	-2.262***	-2.568***
Gender concordance	0.142***	0.041
Exempt	0.360***	0.647***
Age - 75	0.010	0.056***
$Y_{1,t-1}$	0.423***	-0.055**
$Y_{2,t-1}$	-0.739***	-0.046
Factor loading	0.564***	0.574***
No. of obs.	4,992	7,601
No. of patients	1,717	2,579

Table 3: Estimated effects from the dynamic factor model by gender (standard errors based on the delta method; * significant at 10%; ** significant at 5%; *** significant at 1%)

		Men			Women		
		Direct	Indirect	Total	Direct	Indirect	Total
	Y_{1t}						
$Y_{1,t-1}$		0.69***	-0.00	0.68***	0.72***	-0.00*	0.72***
$Y_{2,t-1}$	$0 \rightarrow 1$	-0.05	0.02	-0.03	-0.04*	0.02*	-0.02
Age		-0.00	0.00	-0.00	0.00*	-0.00*	0.00
	$\Pr[Y_{2t} = 1]$ (%)						
$Y_{1,t-1}$	PE at ln(1)	5.8***	-5.1***	1.7***	1.7***	-0.8***	1.2***
$Y_{2,t-1}$	$0 \rightarrow 1$	-7.5***	30.9***	18.0***	1.7***	10.7***	14.3***
Age	PE at 75	-0.2	0.8***	0.6***	0.6***	-0.2***	0.4***
	$\Pr[Y_{3t} = 1]$ (%)						
$Y_{1,t-1}$	PE at ln(1)	0.8***	-0.8***	0.5***	-0.1*	-0.1***	-0.2***
$Y_{2,t-1}$	$0 \rightarrow 1$	-1.0***	8.8***	1.0***	-0.1	0.8***	0.7***
Age	PE at 75	0.0	0.1***	0.1***	0.1***	-0.0***	0.1***

Figure 1: Distribution of patients by gender and year of birth

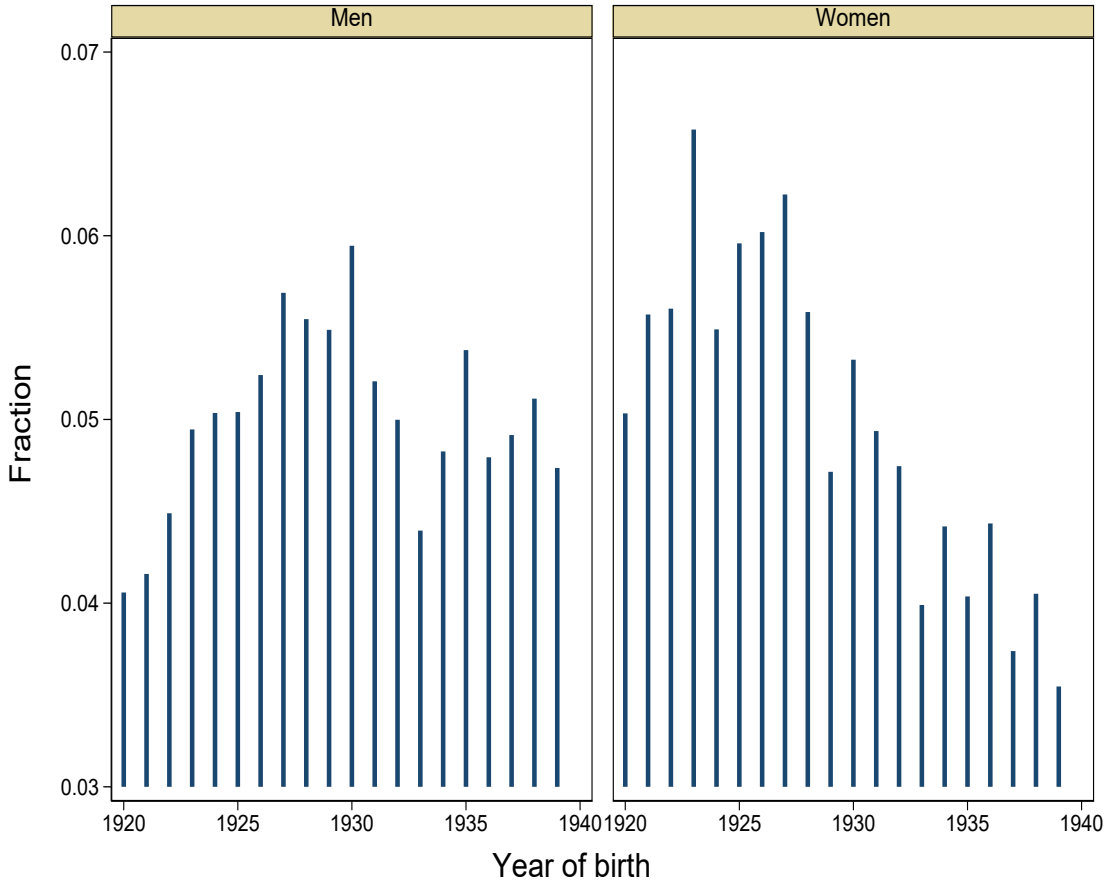


Figure 2: Medical drug use by gender

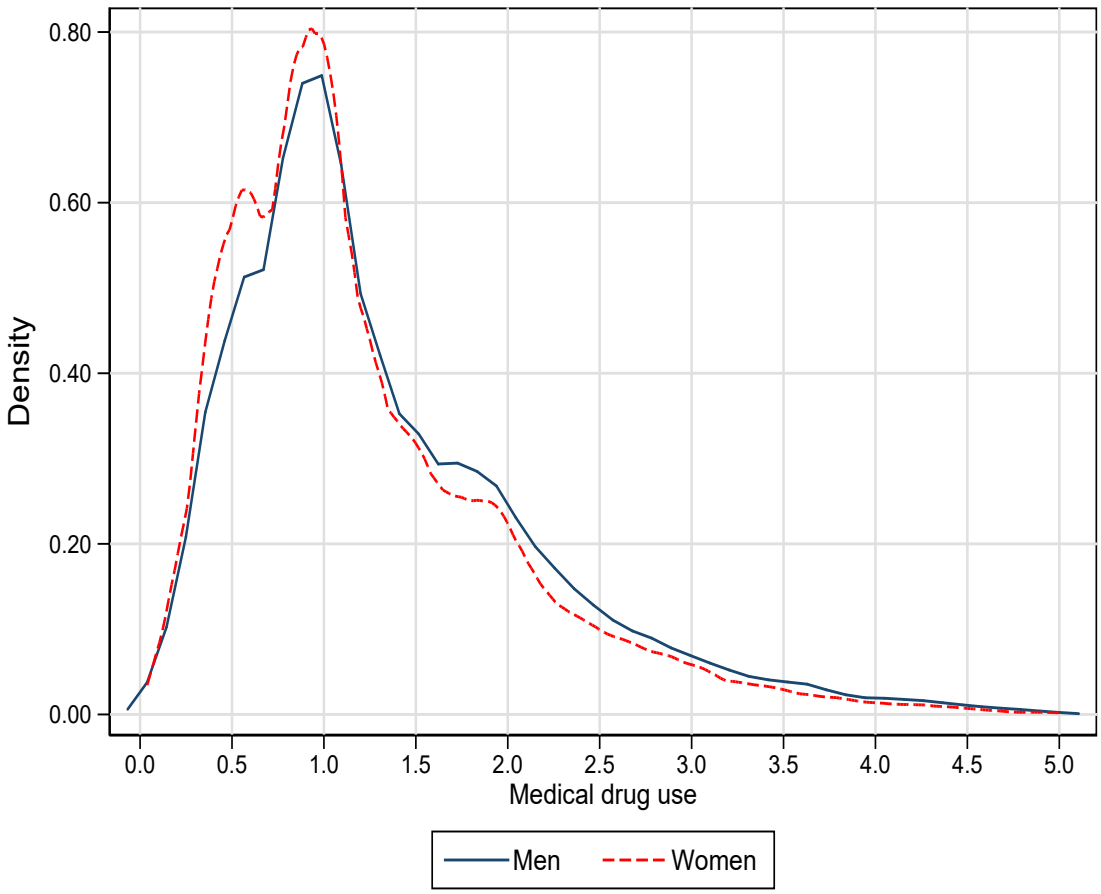
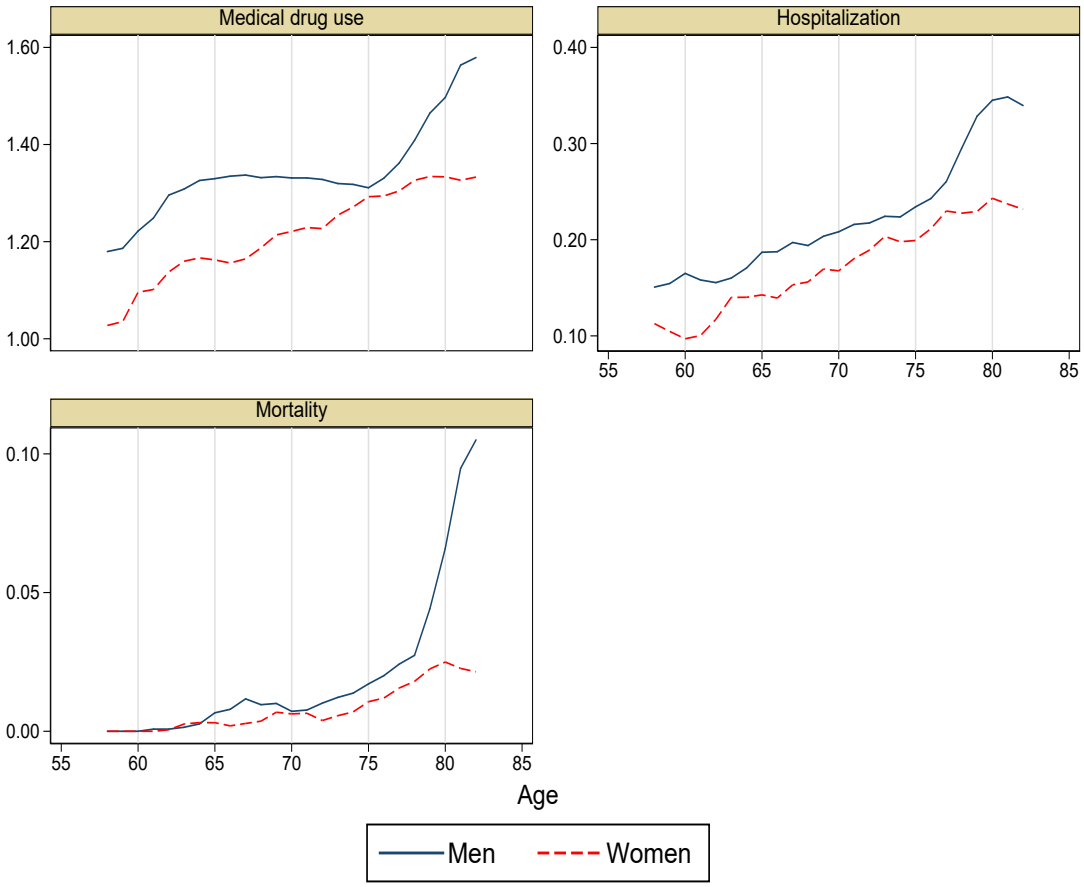


Figure 3: Age profile of average drug use, hospitalization and mortality by gender



Appendix

A Structure of the function g and its Jacobian matrix

Write the vector of p structural parameters in the model as $\psi = (\psi_1, \psi_2)$, where ψ_1 is the subvector of ψ containing the p_1 parameters entering the function g linearly, and ψ_2 is the subvector of ψ containing the p_2 parameters entering g non-linearly, with $p_1 + p_2 = p$. By convention, let ψ_1 be the p_1-subvector of ψ containing the parameters of the low of motion of the outcomes equation (2), andψ_2 be the p_2-subvector of ψ containing the parameters of the reduced-form equation (3) and of the initial condition equation (5). Then, the following relationship links the reduced-form parameters in θ to the structural parameters in ψ

$$\theta = g(\psi) = Q(\psi_2)\,\psi_1,$$

where $Q(\psi_2)$ is a $q \times p_1$ matrix that does not depend on ψ_1. The $p \times q$ Jacobian matrix of $g(\psi)$ is then

$$G(\psi) = \frac{\partial g(\psi)}{\partial \psi} = \begin{bmatrix} \frac{\partial g(\psi)}{\partial \psi_1} \\ \frac{\partial g(\psi)}{\partial \psi_2} \end{bmatrix} = \begin{bmatrix} Q(\psi_2)^\top \\ \psi_1^\top Q_1(\psi_2)^\top \\ \vdots \\ \psi_1^\top Q_h(\psi_2)^\top, \\ \vdots \end{bmatrix}.$$

where $Q_h(\psi_2)$ is the $q \times p_1$ matrix of the partial derivatives of $Q(\psi_2)$ with respect to the h-th element of ψ_2.

To illustrate, consider the special case of three time periods ($T = 2$), one observed outcome Y ($J = 1$), one exogenous time-varying regressor X ($k = 1$), one exogenous time-invariant regressor W and a constant term ($l = 2$). Initial conditions include the initial value of the observed outcome Y, the exogenous time-invariant regressor W, and no pre-sample health shock. In this case, the low of motion of the outcome is

$$Y_{it}^* = a\eta_{it} + b_0 + b_w W_i + b_x X_{it} + b_y Y_{i,t-1} + U_{it}.$$

The law of motion for the latent factor is

$$\eta_{it} = \rho\eta_{i,t-1} + c_x X_{it} + c_y Y_{i,t-1} + \omega V_{it}.$$

The initial condition equation is

$$\eta_{i0} = d_w W_i + d_y Y_{i0}.$$

The reduced-form equation is then

$$Y_{it}^* = b_0^* + b_{wt}^* W_i + \sum_{s=0}^{1} b_{xs}^* X_{i,t-s} + \sum_{s=0}^{1} b_{ys}^* Y_{i,t-s-1} + d_{yt}^* Y_{i0} + \sum_{s=0}^{1} r_s^* V_{i,t-s} + U_{it}, \qquad t = 1, 2.$$

Thus, in this case the vector of $q = 11$ reduced-form parameters is

$$\theta = (b_0^*, b_{w1}^*, b_{w2}^*, b_{x0}^*, b_{x1}^*, b_{y0}^*, b_{y1}^*, d_{y2}^*, d_{y2}^*, r_0^*, r_1^*).$$

Let $\psi = (\psi_1, \psi_2)$ be the vector of $p = 10$ structural parameters, where $\psi_1 = (b_0, b_w, b_x, b_y, a)$ and $\psi_2 = (d_w, d_y, c_x, c_y, \rho)$. In this case, the relationship between θ and ψ can be rewritten as $\theta = g(\psi) = Q(\psi_2)\,\psi_1$, where

$$Q(\psi_2) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \rho d_w \\ 0 & 1 & 0 & 0 & \rho^2 d_w \\ 0 & 0 & 1 & 0 & c_x \\ 0 & 0 & 0 & 0 & \rho c_x \\ 0 & 0 & 0 & 1 & c_y \\ 0 & 0 & 0 & 0 & \rho c_y \\ 0 & 0 & 0 & 0 & \rho d_y \\ 0 & 0 & 0 & 0 & \rho^2 d_y \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \rho \end{bmatrix}.$$

The 10×11 Jacobian matrix of $g(\psi)$ is then

$$G(\psi) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho d_w & \rho^2 d_w & c_x & \rho c_x & c_y & \rho c_y & \rho d_y & \rho^2 d_y & 1 & \rho \\ 0 & \rho a & \rho^2 a & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho a & \rho^2 a & 0 & 0 \\ 0 & 0 & 0 & a & \rho a & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a & \rho a & 0 & 0 & 0 & 0 \\ 0 & d_w a & 2\rho d_w a & 0 & c_x a & 0 & c_y a & d_y a & 2\rho d_y a & 0 & a \end{bmatrix}.$$

Notice that the matrix $G(\psi)$ is of full row rank only if $\rho \neq 0$ and $a \neq 0$.

B Reduced-form estimates

This appendix presents the estimated coefficients of the reduced-form parameters by gender (* significant at 10%, ** significant at 5%, *** significant at 1%.). The MSL estimates of reduced-form parameters are based on 50 draws, using Halton sequences, from the multivariate standard Gaussian distribution of unobserved health shocks and the random pre-determined shock.

	Men			Women		
	Y_1	Y_2	Y_3	Y_1	Y_2	Y_3
γ_1	-0.010	-0.073	0.742***	-0.012	-0.832***	-0.959***
γ_2	-0.013	-0.521***	-0.411***	-0.017	-0.879***	-0.881***
γ_3	-0.009	-0.212	0.098	-0.007	-0.165***	-0.351***
Y_{101}	0.165***	0.058	0.315***	0.147***	0.063*	0.044
Y_{102}	0.044**	-0.014	0.144*	0.039**	0.056	0.200**
Y_{103}	0.040*	-0.033	0.190**	0.030*	-0.105***	-0.210***
Y_{201}	-0.003	0.528***	0.457***	0.049*	0.803***	0.513***
Y_{202}	-0.031	0.302***	0.289***	0.001	0.503***	0.203*
Y_{203}	-0.001	0.362***	0.035	0.024	0.271***	0.243**
Same gender$_1$	0.017	-0.020	0.006	-0.008	-0.104	-0.087
Same gender$_2$	0.001	0.088	0.212*	0.009	0.047	-0.276*
Same gender$_3$	-0.071***	-0.271***	0.241**	0.026	0.137**	0.232**
Exempt$_1$	0.118	0.460***	0.939***	-0.021	0.751***	1.006***
Exempt$_2$	0.061	0.451***	-0.555*	-0.093	0.262	0.972***
Exempt$_3$	0.116	0.705***	0.559***	-0.047	0.451***	1.337***
(Age - 75)$_1$	0.001	0.022***	0.099***	0.000	0.035***	0.078***
(Age - 75)$_2$	0.003	0.018***	-0.067***	0.001	-0.013**	-0.012
(Age - 75)$_3$	-0.006**	0.006	0.063***	-0.002	0.004	-0.047***
$Y_{1t-1,1}$	0.570***	0.141***	0.298***	0.580***	0.067**	0.343***
$Y_{1t-1,2}$	0.209***	-0.132***	0.021	0.203***	-0.009	-0.905***
$Y_{1t-1,3}$	0.022	0.040	0.017	0.045**	0.035	0.352***
$Y_{2t-1,1}$	0.026	0.768***	0.829***	0.015	0.798***	0.484***
$Y_{2t-1,2}$	0.009	0.170***	0.071	0.007	0.524***	0.343***
$Y_{2t-1,3}$	0.040	0.349***	0.349***	-0.062**	0.111**	0.175*
a_1	0.022***	0.850***	0.732***	0.020***	0.826***	0.437***
a_2	0.003	-0.030	-0.322***	0.011	0.412***	0.561***
a_3	0.013	0.261***	0.429***	0.014	0.280***	0.178*
Constant	0.106***	-1.178***	-3.138***	0.083***	-1.634***	-3.357***
Observations		4,992			7,601	
Patients		1,717			2,579	
Sim. log lik.		-32,254.2			-46,572.8	