Journal of Forecasting, *J. Forecast.* **36**, 1–15 (2017)
Published online 21 March 2016 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/for.2409

Forecasting with Micro Panels: The Case of Health Care Costs

DENZIL G. FIEBIG[1]* AND MELIYANNI JOHAR[2]
[1] *School of Economics, University of New South Wales, Sydney, Australia*
[2] *Economics Discipline Group, University of Technology Sydney, Australia*

ABSTRACT

Micro panels characterized by large numbers of individuals observed over a short time period provide a rich source of information, but as yet there is only limited experience in using such data for forecasting. Existing simulation evidence supports the use of a fixed-effects approach when forecasting but it is not based on a truly micro panel set-up. In this study, we exploit the linkage of a representative survey of more than 250,000 Australians aged 45 and over to 4 years of hospital, medical and pharmaceutical records. The availability of panel health cost data allows the use of predictors based on fixed-effects estimates designed to guard against possible omitted variable biases associated with unobservable individual specific effects. We demonstrate the preference towards fixed-effects-based predictors is unlikely to hold in many practical situations, including our models of health care costs. Simulation evidence with a micro panel set-up adds support and additional insights to the results obtained in the application. These results are supportive of the use of the ordinary least squares predictor in a wide range of circumstances. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS health expenditure; risk adjustment; panel data; fixed effects; forecasting

INTRODUCTION

In a recent survey on forecasting with panel data, Baltagi (2008) contrasts the rich and extensive forecasting literature using time series data with the relatively modest literature on forecasting using panel data. Even within the area of panel forecasting, there is very little general research on micro panels where N, the number of individuals, is large relative to T, the number of time series observations. Baltagi (2008) references only one relevant paper by Baillie and Baltagi (1999), who consider forecasting in the context of the standard one-way error component model and compare a number of predictors including models estimated by ordinary least squares (OLS) and fixed effects (FE), as well as variants of the best linear unbiased predictor (BLUP) assuming a random-effects specification. Given the relatively good performance of the fixed-effects predictor (FEP), even when the random effects specification is the true generating process and hence when the FEP is not optimal, Baillie and Baltagi (1999) argue that there is a clear preference for the FEP in order to guard against possible omitted variable biases associated with correlations between unobservable individual specific effects and covariates included in the models. This argument, however, is made without supporting evidence in the case of correlated individual effects, but the preference for an FE approach does mirror the conventional wisdom when parameter estimation, rather than forecasting, is the main objective.

Our primary aim is to provide further comparisons of alternative predictors in the context of the motivating example of risk prediction in health economics. Broadly defined, risk adjustment in health means the use of patient-level information to explain variation in health care utilization, costs and health outcomes (Ellis, 2008). The application of risk adjustment is typically for payment purposes, such as payment to competitive health insurance plans or to health providers, and health insurance premium setting. A good risk adjustment prediction model is one that can forecast accurately the resource use of the individual. In particular, we ask whether the superiority of alternative predictors based on a model estimated by FE over one estimated by OLS is in fact replicated when individual effects are correlated with regressors and in the context of micro panels such as the data we use. The key feature of such panels is the availability of a very large number of individuals (more than 250,000 in our data) but only for a limited time dimension (4 years in our data).

The survey of Baltagi (2008) highlights the existing focus on forecasting in the time dimension. In our risk adjustment application, this task is about forecasting the future health costs to τ periods ahead for a given pool of N individuals where their past cost histories over T time periods are available. For this forecasting task, Baillie and Baltagi (1999) argue that accounting for individual effects has two potential means of impacting the predictor's performance: (i) it may lead to better estimates of the parameters of the explanatory variables; and (ii) in prediction one may explicitly account for the individual effects in constructing the predictor. For (i), FE relies on within-individual variation. Now, in micro panels, this variation is possibly modest. In addition, with small T, the estimates

*Correspondence to: Denzil G. Fiebig, School of Economics, University of New South Wales, Sydney, NSW 2052, Australia.
E-mail: d.fiebig@unsw.edu.au

of the individual effects are likely to be extremely noisy. Both effects indicate that the preference for FEP over OLSP may be threatened in micro panels.

With panel data there is a second type of forecast that is possible where the task is to predict outside the sample to a new group of n individuals for whom there are no past data (e.g. potential customers). In our analysis we consider both forms of forecasting: to distinguish them we refer to forecasting in the cross-section dimension as 'out-of-sample' forecasting, and forecasting in the time series dimension as 'post-sample' forecasting. In general, both types of forecasting are likely to be important and this is especially so in our motivating example of risk adjustment. However, in out-of-sample forecasting, the FEP will not be operational because there will be no estimated individual effects for the new group of n individuals. So in this case the comparison is between the OLSP and alternative FE-based predictors. Following Baillie and Baltagi (1999), we consider a truncated fixed-effect predictor (TFEP) that utilizes FE estimates of the parameters of the explanatory variables but by necessity does not include estimated individual effects in constructing the predictor. In such situations, it is natural to consider the role of time-invariant explanatory variables, which we explore using FE predictors based on the Hausman and Taylor (1981) approach.

There is no shortage of recent research dealing with various aspects of modelling individual health care treatment costs and expenditures. Typical key features of health expenditure data that make modelling a challenge are the presence of a substantial proportion of zero observations (non-users of health services) and positive costs that are highly skewed to the right with long, thick, right-hand tails. Thus many papers compare alternative modelling approaches tailored to accommodate these features using cross-sectional data; see, for example, Jones *et al.* (2014), who ignore the non-users in search for an accurate model of highly skewed data using over 6 million health care observations.

What are in relatively short supply are analyses involving panel data. Two recent survey papers on econometric modelling in health economics serve to illustrate this divide. Jones (2011) surveys the health care cost modelling literature as characterized above but references Jones (2009) for discussion of panel data methods. However, there is only a brief mention of health care cost modelling for risk adjustment in Jones (2009). Instead the applications focused on policy evaluation. Now the choices made on coverage in these two survey papers could merely be driven by the need to narrow the focus, but a close look at the literature does suggest that there has been little work on modelling health care costs using panel data.

Studies where panel have been used to estimate health expenditure models include Seshamani and Gray (2004a,2004b), Stearns and Norton (2004), Albouy *et al.* (2010) and Hill and Miller (2011).[1] In estimating the impact of age and time of death on hospital costs, Seshamani and Gray (2004a) find that ignoring individual fixed effects results in significant omitted variable bias. Their sample is large in both N and T dimensions: over 90,000 individuals with up to 24 years of observations (unbalanced panel). However, because the impact of time-invariant factors such as sex is of key interest, they rely on a random-effects model. Seshamani and Gray (2004b) use the same dataset but have smaller N (9371) and account for the panel nature of the data only through robust standard errors. Stearns and Norton (2004) estimate models using FE but their final predictions of future health costs are made based on a random-effects specification. Albouy *et al.* (2010) consider state dependence in panel data models. Their sample has moderate size N (about 7000) and T (up to 6 years). Instead of fixed effects, the panel structure is accommodated via Wooldridge (2005) type corrections for initial conditions. Hill and Miller (2011) use the US Medical Expenditure Panel Survey (MEPS) data to compare the performance of various models of health expenditure. Although the MEPS data are panel in nature due to quarterly interviews, the study focuses on annual expenditure, so the analysis is cross-sectional.

Thus there is little research that specifically addresses forecasting issues when modelling health costs and expenditures using panel data. Before proceeding to an extensive analysis of our particular micro panel of health care costs, a Monte Carlo study based on Baillie and Baltagi (1999) is conducted in order to illustrate the key issues and trade-offs involved in selecting appropriate predictors.

ECONOMETRIC FRAMEWORK

We abstract from econometric issues that have been the primary subject of many studies comparing approaches for modelling health costs such as transforming the dependent variable, which invites the problems of re-transformation as policymakers require forecasts in raw scale, and accounting for the presence of zero observations. In part, this is to focus attention on issues arising when panel data are available, but it is also a choice supported by past comparisons where simple linear models estimated by OLS do relatively well. As Jones (2011, p. 649) concludes:

> It is notable that the simple linear model, estimated by OLS, performs quite well across all of the criteria, a finding that has been reinforced for larger datasets than the one used here.

[1]There are also several descriptive studies using commercial MarketScan Databases in the USA, which contains data from the employer and health plan sources concerning medical and drug data for several million commercially insured individuals, including employees, their spouses and dependants collected since 1995 (e.g. Aizcorbe *et al.*, 2012) and MEPS data (e.g. Zuvekas and Olin, 2009; Bernard *et al.*, 2011).

This is in fact what we have found with preliminary analyses using our data on health expenditures (see Ellis *et al.*, 2013). Our expenditure data happen to have a small proportion of zero observations due to the setting of a universal public health care system. Thus the forecasting comparison here is kept narrowly focused on the differences between predictors using estimates produced by OLS and by variants using FE estimates in the context of a basic linear panel model with a common set of regressors.

The model is represented by

$$y_{it} = \alpha + x_{it}^{'}\beta + z_i^{'}\gamma + \mu_i + v_{it}; i = 1, \ldots, N, N+1, \ldots, N+n; t = 1, \ldots, T, T+1, \ldots, T+\tau \quad (1)$$

where there are NT within-sample observations and, depending on the forecasting task, an extra n individuals or an extra τ time periods. In using equation 1 to extrapolate beyond the data used for estimation purposes, one could consider producing forecasts for $y_{i,T+1}, \ldots, y_{i,T+\tau}$. In other words, one could forecast future costs for the $i = 1, \ldots, N$ sample of individuals used for estimation of the model parameters; forecasting in the time dimension will be termed *post-sample*. But it would also be relevant to consider a different sample of individuals not used in the estimation stage and predict their costs at a given time t; forecasting $y_{N+1,t}, \ldots, y_{N+n,t}$ in the cross-sectional dimension will be termed *out-of-sample*. (Obviously there is also the case of post-out-of-sample forecasting, $y_{N+n,T+\tau}$). Several alternative predictors are considered, although not all will be available for these different forecasting tasks. In this set-up the distinction between regressors that vary over both time and individuals (x_{it}) and those that are time-invariant (z_i) has been made explicit and we have allowed for the potential presence of unobservable time-invariant factors, μ_i.

In their simulation experiments, Baillie and Baltagi (1999) specify the data-generating process as a classical one-way error component model, where the μ_i in equation 1 are assumed random and independent of all regressors. In their comparison of predictors they emphasized the impact of accounting for individual effects on both estimation and prediction. But because of their choice of data-generating process, the resulting estimation problem abstracts from possible biases in coefficient estimates and concentrates on relative efficiencies of alternative predictors. While this represents a reasonable base case, the alternative situation where the μ_i are potentially correlated with the regressors is an important and practically relevant extension. Here the generation of consistent parameter estimates becomes an issue in comparing alternative predictors. The Baillie and Baltagi (1999) framework is also extended by considering the out-of-sample forecasting task.

OLS applied to the pooled data, ignoring the individual specific effects, yields parameter estimates denoted by $\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma}$ and produces the OLS predictor (OLSP), given by

$$\text{OLSP}: \ \widetilde{y}_{it} = \widetilde{\alpha} + x_{it}^{'}\widetilde{\beta} + z_i^{'}\widetilde{\gamma}$$

For estimation by FE, the intercept is parametrized to be the mean of the individual specific effects, implying that the estimated μ_i are restricted to have a zero mean. This provides options in defining FE-based predictors. If the estimated μ_i are not used in forming forecasts then we call this the truncated FE predictor and denote it by TFEP. Alternatively, adding in the estimated fixed effects yields the FE predictor denoted by FEP. Thus, denoting the FE parameter estimates by $\hat{\alpha}, \hat{\beta}, \hat{\mu}_i$, these alternative predictors are defined as

$$\text{TFEP}: \ \hat{y}_{it} = \hat{\alpha} + x_{it}^{'}\hat{\beta}$$
$$\text{FEP}: \ \hat{y}_{it} = \widetilde{\alpha} + x_{it}^{'}\hat{\beta} + \hat{\mu}_i$$

Consideration of both types of FE predictors is in part motivated by Baillie and Baltagi (1999), who distinguished between the impact of accounting for individual effects in parameter estimation and then in forming the predictor. Here, though, it also derives from the feasibility of approaches. In a post-sample forecasting task, both of these FE approaches are feasible predictors but only TFEP is available in the out-of-sample forecasting task because of the unavailability of estimated μ_i for the new sample of n individuals.

Note that the time-invariant z's will appear in the OLSP but not in either of the FE predictors. From a forecasting perspective this may or may not be an important source of differentiation between predictors. In post-sample forecasting, FEP provides a very flexible alternative to allowing for the z's in the predictor. However, in the out-of-sample forecasting task one might expect that OLSP potentially has an advantage over TFEP because of the inclusion of the z's in the predictor.

In the case of out-of-sample forecasting an additional FE-based predictor is defined by generating estimates of from the following regression model:

$$\hat{\mu}_i = \theta + z_i^{'}\gamma + \omega_i \quad (2)$$

where the estimated individual specific effects are regressed on the z's. This estimator is discussed in Hsiao (1986) and is a Hausman and Taylor (1981) type estimator that would result under the assumption that any correlation

between unobservables and regressors is confined to the time-varying regressors. These two-step estimates denoted by $\hat{\theta}, \hat{\gamma}$ are then used in conjunction with the first-step FE estimates to form the Hausman and Taylor predictor HTP:

$$\text{HTP}: \quad \hat{y}_{it} = \hat{\alpha} + x_{it}'\hat{\beta} + \left(\hat{\theta} + z_i'\hat{\gamma}\right) = \left(\hat{\alpha} + \hat{\theta}\right) + x_{it}'\hat{\beta} + z_i'\hat{\gamma}$$

Using our data we compare the performance of these alternative predictors for both post-sample and out-of-sample forecasting tasks. In addition, we distinguish concurrent and prospective specifications. Equation 1 depicts a concurrent specification where the current period's expenditure is explained by current period covariates. Alternatively, the prospective model has next year's total health expenditure as the dependent variable and current period covariates. In practice, the prospective specification is preferred for budgeting as it indicates what the health care costs would be in the future. However, in terms of model fit, the concurrent specification has a much better fit than the prospective specification. By analysing both models, we can therefore check the sensitivity of our results to model fit. Performance is evaluated in terms of forecast mean squared errors (MSE) with the ranking based on minimizing MSE (mean absolute prediction error (MAPE) was also calculated but leads to qualitatively the same results). We also compute the predictive ratio, which is a group-level measure of predictive accuracy. It involves adding up the total predicted expenditure for a group of individuals and comparing that value to the actual expenditure for the same group. A predictive ratio that is closer to 1 indicates a better fit.

In the first instance, a Monte Carlo study is conducted in order to illustrate the key issues and trade-offs involved in selecting the appropriate predictor. The Baillie and Baltagi (1999) experimental design is used as a base and extensions are restricted to cases where FE approaches—FEP and TFEP—produce consistent parameter estimates. Despite this restriction, it is not a priori obvious that an FE-based predictor is necessarily always superior to the OLSP. In Baillie and Baltagi (1999) the often substantial superiority of FEP over OLSP derives from the individual heterogeneity in outcomes that is accommodated by estimating individual effects. When forecasting out-of-sample, this adjustment is not available and hence puts the feasible FE predictor, TFEP, back on a more equal footing with the OLSP. Also what is required is a good approximation of the conditional mean function and this is not guaranteed by inserting consistent estimates of a subset of parameters, especially when these consistent estimators may have large variances because of limited within-variation. These arguments have prompted the inclusion of a simulation study in order to provide some guidance on what to expect when we undertake our extensive empirical analysis. The situation is further complicated when time-invariant variables are available and HTPs are considered. This additional issue is not addressed in the simulation study but is left to the substantive application that follows, as is the comparison between concurrent and prospective specifications.

MONTE CARLO EXPERIMENT

Simulation design

The initial data-generating process (DGP) to be considered is a variant of equation 1 in which there are no time-invariant explanatory variables and a single time-varying covariate, implying

$$y_{it} = \alpha + x_{it}\beta + \mu_i + v_{it} \tag{3}$$

where $\alpha = 5$, $\beta = 0.5$, $v_{it} \sim$ i.i.d. $N\left(0,\ \sigma_v^2\right)$, and the DGP for x_{it} is given by

$$x_{it} = 0.1t + \ 0.5x_{i,t-1} + \varepsilon_i + \omega_{it} \tag{4}$$

where ω_{it} is uniformly distributed on the interval $[-0.5,\ 0.5]$ and $x_{i0} = 5 + 10\omega_{i0}$, with the first 20 observations discarded. The unobserved individual effects are correlated with the explanatory variable through the following specification:

$$\mu_i = \varepsilon_i + \eta_i \tag{5}$$

with $\varepsilon_i \sim$ i.i.d. $N\left(0,\ \sigma_\varepsilon^2\right)$ and $\eta_i \sim$ i.i.d. $N\left(0,\ \sigma_\eta^2\right)$, which implies $\text{var}(\mu_i) \equiv \sigma_\mu^2 = \sigma_\varepsilon^2 + \sigma_\eta^2$ and $\text{cov}(x_{it},\ \mu_i) = \sigma_\varepsilon^2$. This reduces to the Baillie and Baltagi (1999) experimental design when $\sigma_\varepsilon^2 = 0$. In order to facilitate comparisons, their results are reproduced in what is termed Experiment 1. This involves varying $\rho = \sigma_\mu^2 \,/\left(\sigma_\mu^2 + \sigma_v^2\right)$ as 0, 0.3, 0.6 or 0.9 with $\sigma_\mu^2 + \sigma_v^2$ fixed at 20 over 1000 replications for each design point. In addition to one-period-ahead post-sample forecasts, we also provide out-of-sample forecasts for a holdout sample of $n = 50$ individuals. Experiment 2 then repeats the analysis, adding correlation between the individual effects and the explanatory variable induced by setting $\sigma_\varepsilon^2 = 0.81$, a value that implies correlations between 0.10 and 0.25. All comparisons are done in terms of forecast mean squared errors (MSE).

Apart from introducing correlations between unobserved individual effects and the explanatory variable, the set-up of Baillie and Baltagi (1999) needed to be extended in other respects to better reflect the type of applications we seek to explore. In their experiments they set $N=50$ or 500 and $T=10$ or 20, while in Experiments 1 and 2 we specify $N=500$ or 1000 and $T=3$, 10 or 20. To get even closer to a more realistic situation, Experiment 3 is developed based on the actual data to be used in our application. Using total health expenditures ($\exp_{it}$) as the dependent variable and number of standard (less than 20 minutes) visits to a general practitioner (GP) in a year (gp_{it}) as the sole explanatory variable, a fixed-effects specification is estimated for $N=1000$ and $T=3$ and the resulting parameter estimates are taken to be the 'truth' to generate data for two panel configurations ($N=500$ or 1000 and $T=3$). The resulting DGP is given by

$$exp_{it} = 2.198 + 0.279\mathrm{gp}_{it} + \mu_i + v_{it} \tag{6}$$

where $\sigma_\mu^2 = 46.06$ and $\sigma_v^2 = 41.65$ and the gp_{it} values are fixed in this 'real data' experimental design. Note that the estimated individual effects are constrained to sum to zero, implying that the intercept represents the overall mean of total health expenditures (in thousands). Further details on the data will be provided in the 'Data' section, below.

The real data design is completed by specifying a source of omitted variable bias by generating the μ_i as follows:

$$\mu_i = \lambda \mathrm{sp}_i + \eta_i \tag{7}$$

where sp_i is the mean number of initial specialist consultations in a year for each individual from our data normalized to have a unitary variance, again taken to be fixed over replications, and $\eta_i \sim i.i.d.\ N\left(0,\ \sigma_\eta^2\right)$. Setting $\lambda=0$ and varying $\rho = \sigma_\mu^2 / \left(\sigma_\mu^2 + \sigma_v^2\right)$ while keeping fixed $\sigma_\mu^2 + \sigma_v^2 = 87.71$, mimics Experiment 1, but with a very different explanatory variable. Compared to the Baillie and Baltagi (1999) design, the contribution to total variation in the dependent variable due to the variation in the unobservables is much smaller here, as is the within-variation relative to the between-variation in the single explanatory variable. Because FE estimation relies on the within-variation we might expect a deterioration in the relative performance of the FEP in this alternative design.

Setting $\lambda=1$ reproduces the sample correlation of 0.364 between sp_i and gp_{it} and provides a real data version of Experiment 2. While sp_i forms part of the DGP, it is taken to be unobservable by the researcher and thus is not part of the model specification estimated by OLS and FE. Because it is a time-invariant variable, this does not affect the FE estimation but will induce parameter biases when OLS is used and $\lambda \neq 0$.

Simulation results

Table 1 provides the MSE results for Experiment 1 when $\sigma_\varepsilon^2 = 0$. The post-sample forecasting analysis for $N=500$ and $T=10$ or 20 reproduces that part of Table 1 of Baillie and Baltagi (1999) corresponding to the predictors of interest here, namely OLSP and FEP. The key feature of this subset of the results is that the FEP dominates OLSP except when $\rho=0$ and the difference increases markedly as ρ increases. Baillie and Baltagi (1999) demonstrate that FEP runs a close second to the operational optimal predictor (they term this the 'ordinary predictor') based on the true random-effects DGP and emphasize the importance of accounting for individual effects in estimation and prediction. Further, they conclude that the FEP is recommended in practice because of this relatively good simulation performance and its robustness to correlation between random effects and regressors. Note that they do not extend their experimental design to provide simulation evidence in support of the robustness claim.

The remaining results in Table 1 provide some initial qualifications to the main findings of Baillie and Baltagi (1999). First, the post-sample forecasts when $T=3$ exhibit a similar pattern except there is a relative deterioration in FEP attributable to the reduced within-variation in the regressor that is associated with the short time dimension. This means that for $\rho=0$ the OLSP dominates FEP by a greater margin than when T is larger. The second feature relates to the out-of-sample forecasting section of the results where the superiority of an FE approach over OLSP is eliminated. Baillie and Baltagi (1999) argue that accounting for individual effects can lead to better estimates of the parameters of the explanatory variables and also can provide estimates of the individual effects for use in constructing the predictor. In out-of-sample forecasting, TFEP accounts for individual effects in estimation, but because forecasting relates to a new sample of individuals there are no estimated individual effects, and only a single overall mean effect can be used for prediction. The result that the performance of the two predictors, TFEP and OLSP, is essentially the same for all N and T combinations serves to emphasize that the importance of accounting for individual effects relates to their use in prediction rather than estimation.

Table 2 provides the MSE results for Experiment 2, where OLS no longer produces consistent parameter estimates because of the non-zero correlation between the explanatory variable and the unobserved time-invariant effects. There are no results for $\rho=0$ because this implies no unobserved time-invariant effects. Setting $\sigma_\varepsilon^2 = 0.81$ implies modest levels of correlation (0.10–0.25) between the explanatory variable and the unobserved effects, but the biases in the OLS parameter estimates are substantial. For example, with $N=1000$, $T=3$ *and* $\rho=0.9$, the mean correlation calculated over the 1000 replications of the experiment was 0.143, while the means of the OLS estimates of $\alpha=5$ and $\beta=0.5$ were 1.452

Table I. Post-sample and out-of-sample mean squared errors: Experiment 1

	Post-sample		Out-of-sample	
	OLSP	FEP	OLSP	TFEP
N= 500, T= 10				
$p=0$	20.05	22.06	20.03	20.03
$p=0.3$	20.00	15.45	19.97	19.97
$p=0.6$	19.97	8.83	19.92	19.92
$p=0.9$	19.95	2.21	19.88	19.88
N= 500, T= 20				
$p=0$	20.00	20.99	20.03	20.03
$p=0.3$	20.00	14.66	20.06	20.06
$p=0.6$	20.00	8.43	19.99	19.99
$p=0.9$	20.00	2.10	20.07	20.07
N= 500, T= 3				
$p=0$	20.03	26.66	20.06	20.08
$p=0.3$	20.01	18.68	20.06	20.07
$p=0.6$	20.02	10.67	20.11	20.10
$p=0.9$	20.02	2.67	20.11	20.10
N= 1000, T= 3				
$p=0$	20.01	26.63	20.15	20.16
$p=0.3$	20.03	18.67	20.05	20.05
$p=0.6$	20.01	10.67	20.16	20.16
$p=0.9$	19.98	2.67	20.29	20.28

Notes: (i) The first two *N*, *T* pairs were chosen to be comparable with the design choices of Baillie and Baltagi (1999), whereas the second two *N*, *T* pairs were chosen to be more representative of micro panels. (ii) OLSP is the OLS predictor; TFEP is the FE predictor without the estimated individual specific effects; FEP is the FE predictor with the estimated individual specific effects. We have checked that these results are not sensitive to the seeds chosen.

Table II. Post-sample and out-of-sample mean squared errors: Experiment 2

	Post-sample		Out-of-sample	
	OLSP	FEP	OLSP	TFEP
N= 500, T= 10				
$p=0.3$	19.86	15.45	19.51	20.05
$p=0.6$	19.76	8.78	19.54	20.07
$p=0.9$	19.89	2.21	19.62	20.05
N= 500, T= 20				
$p=0.3$	20.09	14.71	19.71	20.00
$p=0.6$	20.12	8.41	19.72	19.98
$p=0.9$	20.16	2.09	19.76	20.02
N= 500, T= 3				
$p=0.3$	19.39	18.66	19.47	20.24
$p=0.6$	19.47	10.72	19.35	20.18
$p=0.9$	19.32	2.67	19.47	20.20
N= 1000, T= 3				
$p=0.3$	19.41	18.67	19.47	20.16
$p=0.6$	19.36	10.68	19.44	20.16
$p=0.9$	19.41	2.67	19.41	20.03

Notes: (i) The first two *N*, *T* pairs were chosen to be comparable with the design choices of Baillie and Baltagi (1999), whereas the second two *N*, *T* pairs were chosen to be more representative of micro panels. (ii) OLSP is the OLS predictor; TFEP is the FE predictor without the estimated individual specific effects; FEP is the FE predictor with the estimated individual specific effects.

and 1.347, respectively. Despite these large parameter biases the pattern in the post-sample MSEs was largely unchanged from Experiment 1: the FEP dominates OLSP and the difference increases as ρ increases.

The pattern in the out-of-sample forecasting section of the results has changed though. In Experiment 1 there was essentially no difference between TFEP and OLSP. Now, while the differences are modest, OLSP uniformly dominates TFEP despite the fact that OLS produces severely biased estimates of the true parameter values. This is an important finding if the models are primarily used for out-of-sample forecasting, as in the case of risk adjustment models in health care.

The out-of-sample superiority of OLSP over TFEP is an artefact of a more general phenomenon. Consider a classical linear regression variant of equation 1 where $\mu_i = 0$, implying that the optimal predictor, in a lowest MSE sense, would require replacing the unknown parameters with their OLS estimates. If instead z_i is not observed, giving rise to a standard omitted variable situation, then the 'misspecified' predictor is now only a function of x_{it}. Using OLS to estimate the short regression produces consistent estimates of population parameters that incorporate the biases due to the impact of omitting z_i and corresponds to the best linear predictor of y_{it} conditional on x_{it} alone. With panel data, one can obtain consistent estimates of α and β using an FE estimator, but using these in the prediction equation that is solely a function of x_{it} means that there is no account being taken of z_i for the purposes of prediction and will produce poor forecasts in the case where omitted variable biases occur. It is essentially the same as taking the optimal predictor when both x_{it} and z_i are available in equation 1 but then specifying $\gamma = 0$.

Table 3 provides the MSE results for Experiment 3, where the real data design is utilized. The table is partitioned into two parts; the first part mirrors Experiment 1, where the explanatory variable is uncorrelated with the unobserved time-invariant effects, and the second part, where such correlation is introduced, mirrors Experiment 2. Qualitatively the overall results are similar. Even with a very different explanatory variable, taken from our data, FEP does well in post-sample forecasting irrespective of whether OLS parameter estimates are impacted by omitted variable biases or not. The situation with out-of-sample results has changed somewhat in that OLSP now dominates TFEP in both the uncorrelated and correlated cases. Previously they were almost identical in the uncorrelated case but now the within-variation is relatively small compared to the between-variation and this adversely impacts the performance of the TFEP.

These Monte Carlo results are not meant to represent an exhaustive investigation of issues relevant to choice of predictors with micro panel data. Rather they serve to highlight some important dimensions of the comparison between alternative predictors that were not evident in the work of Baillie and Baltagi (1999), and they provide some indication of what to expect in the case study to follow. In particular, we expect the OLSP to perform relatively better in the out-of sample forecasting task compared to the post-sample forecasting task.

DATA

Our data are derived from the 45 and Up Study of 267,153 New South Wales (NSW) residents linked to several administrative data sources of health costs from 2006 to 2009: hospital inpatient data and emergency department (ED) data (linked by the NSW Centre for Health Record Linkage http://www.cherel.org.au/), Medical Benefits

Table III. Post-sample and out-of-sample mean squared errors: Experiment 3

	Post-sample		Out-of-sample	
	OLSP	FEP	OLSP	TFEP
Uncorrelated				
N= 500, T= 3				
$p=0$	87.47	116.58	88.62	93.84
$p=0.3$	87.70	81.53	87.97	93.19
$p=0.6$	87.61	46.79	88.02	93.31
$p=0.9$	87.20	11.72	87.96	92.75
N= 1000, T= 3				
$p=0$	87.70	116.86	88.08	93.33
$p=0.3$	87.71	81.91	87.56	92.79
$p=0.6$	87.82	46.86	88.38	93.23
$p=0.9$	87.73	11.68	87.94	92.96
Correlated				
N= 500, T= 3				
$p=0.3$	87.25	81.90	87.60	92.10
$p=0.6$	87.55	46.78	88.26	92.70
$p=0.9$	87.39	11.68	89.58	93.20
N= 1000, T= 3				
$p=0.3$	87.45	81.80	87.59	92.24
$p=0.6$	87.37	46.80	87.42	92.09
$p=0.9$	87.34	11.67	87.59	92.06

Notes: (i) Here only the *N*, *T* pairs more representative of micro panels are used, but results are provided when the individual effects are uncorrelated with the explanatory variable (first two panels) and when there is correlation (second two panels). (ii) OLSP is the OLS predictor; TFEP is the FE predictor without the estimated individual specific effects; FEP is the FE predictor with the estimated individual specific effects.

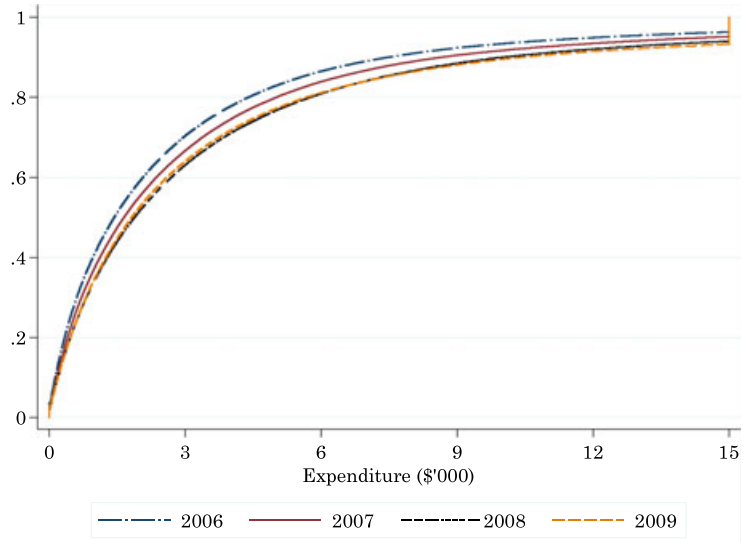
Schedule (MBS) data for medical services such as GP and specialist consultations and Pharmaceutical Benefits Scheme (PBS) data of prescription drugs, for which a government subsidy was paid. The survey was collected only once during this period (45 and Up Study Collaborators, 2008), but the health records of the survey respondents are a panel. We exclude voluntary participants and those respondents with invalid age (0.1%), and those who died during the study period. The final sample is 1,056,096 person-years. The average age of the survey respondents is 63 years.

Individual annual total health expenditure is calculated as the sum of costs of hospital services, charges for MBS items and prices of PBS drugs in any given year. The cost of hospital services is input using the *NSW Costs of Care Standards 2009/10* guidelines released by NSW Department of Health (NSW Health, 2011). For hospitalization, it varies by diagnostic group, type of treating hospital, type of care (e.g. overnight or same day), length of stay, intensive care unit (ICU) hours and the use of ventilation machine. For ED presentations, cost varies with hospital type, urgency category and whether or not the patient is subsequently admitted. All expenditures are annual and indexed to constant 2009 AUD.

A common feature of health expenditure data is positive skewness. Figure 1 illustrates that while half of the population only use less than $1700 worth of health care, the top 25% of the population use more than $4200 worth of health care and the very top 5% use in excess of $9000 worth of health care. This pattern is consistent across all years, suggesting the absence of any relevant structural break in demand within our study period that needs to be accounted for. In the last year, however, we observe higher prevalence of very high expenditures, as indicated by longer vertical lines at the top scale of $15,000.

Another typical feature of health expenditure data is a large mass of zero expenditure; the so-called non-users of health care who may be self-selected. However, our setting is the Australian universal public health system, which ensures access to health services by all. In our data, less than 3% of individuals have zero expenditure in any given year. This could also be explained by our older sample of individuals, whose demand for health services is relatively higher than the general population. The mean expenditure was $3449 in 2006, $4054 in 2007, $4677 in 2008 and $5004 in 2009.

Time-varying regressors (x_{it}) are diagnoses during hospitalization and drug groups. To summarize this rich information into a manageable number of variables to be put into the regression model, we use a US-based risk adjustment tool called DxCG Risk Solutions, developed by Verisk Health. This is a standard approach in the risk adjustment literature. The software, which extends the classification system used by the US Medicare program for paying competing health plans, organizes diagnosis codes (International Statistical Classification of Diseases version 10) and drug codes (Anatomical Therapeutic Chemical Classification) into a large number of non-mutually exclusive categories, and imposes hierarchies on diseases and drugs, so that more serious or expensive conditions take precedence over less serious or expensive conditions. The software also performs a number of data-cleaning steps to identify illegal (e.g. coding errors) or invalid (e.g. male pregnancies) diagnoses. It has been used by numerous academic papers in the US such as Ash *et al.* (2001), Einav *et al.* (2013) and Zhao *et al.* (2005). The result of the grouping is 110 non-mutually exclusive dummy variables for diagnoses (related condition categories, RCCs) and 123 non-mutually exclusive dummy variables for drugs (RX groups). Age and its interaction with gender are also included as time-varying regressors.



Note: y-axis is the cumulative distribution sorted by expenditure. Expenditures above $15,000 are top-coded to make the graph for the bulk of observations more visible.

Figure 1. Health expenditure cumulative distribution function by year.

All other variables from the 45 and Up Survey are time-invariant, some because they are only measured at one point in time. To explore the role of various observed individual specific factors in explaining variation in health expenditure, we define three sets of time-invariant variables. The first set z_{1i} consists of basic demographic characteristics, such as sex, marital status, education, the possession of a health care concession card, region, foreign born, language status and skin colour to capture ethnicity. The second set, z_{2i}, augments z_{1i} by including self-reported health variables such as self-assessed general health and major chronic illnesses, such as diabetes, hypertension, cancers, heart disease, broken bones and asthma. The third set, z_{3i}, augments z_{2i} by including socioeconomic characteristics, such as income, employment and private health insurance status, and lifestyle variables such as smoking, obesity and alcohol consumption. The last set of additional variables is potentially endogenous.

FORECASTING RESULTS

All of the analyses use data drawn from a balanced panel of 264,024 individuals observed for 4 years. For the prospective specification only 3 years of data are available. These data are then divided into estimation and holdout samples that vary depending on the type of forecasting analysis being undertaken. Summary statistics of selected variables are provided in the Appendix.

In the first set of results provided in Table 4 we consider post-sample prediction. Here the holdout sample comprises 1 year of data for all individuals. Predictive MSEs are reported for both concurrent and prospective specifications, as are comparable within-sample measures as a baseline. Within sample, and for both concurrent and prospective specifications, the ranking of predictors is the same with FEP, being better than OLSP, which in turn is better than TFEP. What is noteworthy is that the superiority of OLSP over TFEP is more pronounced in the prospective specification. What is also clear is the superiority of the concurrent specification in terms of fit over the prospective, which is not unexpected because the former captures contemporaneous associations.

Setting aside 1 or 2 years of data for post-sample prediction, the superiority of FEP disappears. In all cases and for both the concurrent and prospective specifications, OLSP produces the lowest predictive MSE. For the concurrent specification the differences are relatively small across all three predictors, but FEP is not necessarily superior to TFEP. For the prospective case the post-sample ordering is clearer, with the use of estimated individual fixed effects in FEP improving prediction performance relative to TFEP although still remaining inferior to the OLSP.

Note that, unlike the earlier simulation results, the predictive models include both time-varying and time-invariant predictors. These will remain in the OLSP and one would expect this feature to assist predictive performance relative to TFEP. While the time-invariant predictors will also be absent from FEP, one expects the use of estimated individual fixed effects, which are so important in producing a good within-sample fit, to compensate. For our application this proves not to be the case in the post-sample results presented in Table 4.

Table 5 provides the analysis of out-of-sample predictions. Here the holdout sample refers to a randomly selected 20% sample of 52,932 individuals. We also experimented with alternative holdout samples and find similar patterns. In terms of both within-sample fit and out-of-sample predictions and in both concurrent and prospective

Table IV. Predictive mean squared error comparison of alternative predictors: post-sample

	Concurrent MSE	Prospective MSE
Within sample fit		
OLSP	30.24	76.19
TFEP	31.04	104.7
FEP	18.78	40.58
Prediction, $T = 3$, $\tau = 1$		
OLSP	48.68	–
TFEP	49.81	–
FEP	49.65	–
Prediction $T = 2$, $\tau = 1$		
OLSP	30.94	106.1
TFEP	32.59	158.0
FEP	33.20	122.2
Prediction $T = 2$, $\tau = 2$		
OLSP	49.76	–
TFEP	51.30	–
FEP	54.60	–

Note: $N = 264{,}024$ in all cases. For within-sample fit $T = 3$ for prospective and $T = 4$ for concurrent. τ -values indicate whether the prediction is one or two periods ahead. The covariates of OLSP are the full set of $z3$, RCC dummies and RX dummies.

Table V. Predictive mean squared error comparison of alternative predictors: out-of-sample

	Concurrent MSE	Prospective MSE
Within-sample fit		
OLSP($z1$)	29.55	74.92
OLSP($z2$)	29.53	74.36
OLSP($z3$)	29.51	74.20
TFEP	30.26	102.6
FEP	18.19	39.19
HTP ($z1$)	30.15	96.52
HTP($z2$)	30.11	90.66
HTP($z3$)	30.09	89.85
Prediction		
OLSP($z1$)	33.37	85.23
OLSP($z2$)	33.34	84.68
OLSP($z3$)	33.33	84.54
TFEP	34.12	113.7
HTP($z1$)	34.02	107.4
HTP($z2$)	33.99	101.5
HTP($z3$)	33.97	100.7

Notes: (i) In all cases $T = 3$ for the prospective and $T = 4$ for the concurrent specifications. $N = 211{,}728$ for the within-sample fit, then models are used to predict for a 20% holdout sample of $n = 52{,}932$. Because N is different from that used to produce Table 4, the comparable within-sample MSEs will be different. (ii) Three variants of the OLS predictor (OLSP) and the Hausman and Taylor predictor (HTP) are defined depending on which of three sets of time-invariant explanatory variables are used. Moving from $z1$ to $z3$ (and hence from OLSP($z1$) and HTP($z1$) to OLSP($z3$) and HTP($z3$)), more time-invariant regressors are added, but at the same time the exogeneity of these additional regressors becomes more problematic.

specifications, it makes no substantive difference to the performance of OLSP which z's are used. Recall that moving from $z1$ to $z3$, more time-invariant regressors are added but at the same time the exogeneity of these additional regressors becomes more problematic. While FEP delivers the best within-sample fit, it is not feasible for out-of-sample forecasting; individual effects can only be estimated for the N individuals in the estimation sample. One might then expect the augmentation with the z's will add significantly to performance of FE based predictors, but Table 5 shows that the gain is more modest. Within-sample fit and out-of-sample predictive performance of HTP changes little with choice of z's for the concurrent specification but does deliver modest improvement in the prospective specification as more z's are added. OLSP is the best-performing predictor although the differences are relatively small for the concurrent specification, whereas they are substantial in the prospective specification.

Thus far, predictive performance has been gauged on the basis of predictive MSEs, but the relative performance of alternative predictors may very well be sensitive to choice of metric. As an alternative, selected comparisons are repeated using the predictive ratio, which is commonly used in the risk adjustment literature. Unlike predictive MSE, which measures model performance at the level of the individual, risk ratios measure performance at a group level. For each decile of actual expenditure, we calculate the ratio of the aggregate actual expenditure to the aggregate predicted expenditure. These ratios are represented in Figures 2–4 and predictor superiority is judged by how close these ratios are to unity.

Figure 2 provides a comparison of predictive performance post sample for both concurrent and prospective specifications, together with within-sample performance as a benchmark. In both concurrent and prospective specifications, overestimation of expenditures is the norm, and risk ratios are almost always less than unity for low-risk (expenditure) deciles 1–9. The extent of overestimation is often extreme and this is more so for the prospective case. Expenditure in the highest decile is underestimated and again this is much more pronounced for the prospective specification. The ranking of predictors is stable across specifications. The FEP dominates OLSP both within sample and post sample.

Turning to the out-of-sample results, the concurrent and prospective specifications are provided separately in Figures 3 and 4. Because the choice of time-invariant regressors had limited impact on the performance of alternative predictors, all time-invariant regressors are included in OLSP and HTP. For both the concurrent and prospective specifications, there is a recurrence of the pattern observed in Figure 2, where expenditure in all deciles except the top decile are typically overestimated and the expenditures in the largest decile are underestimated. Also, FEP dominates within sample but is not operational for out-of-sample forecasts.

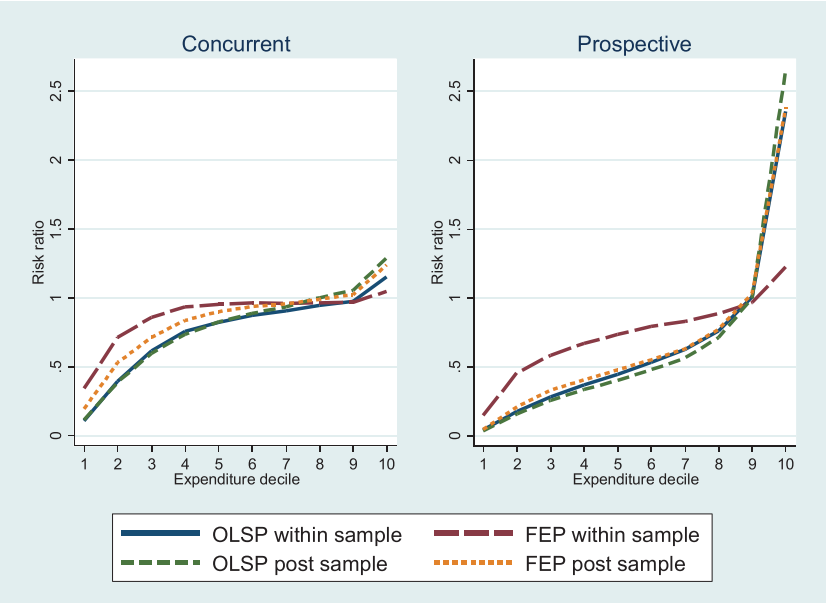


Figure 2. Predictive risk ratios for post-sample prediction

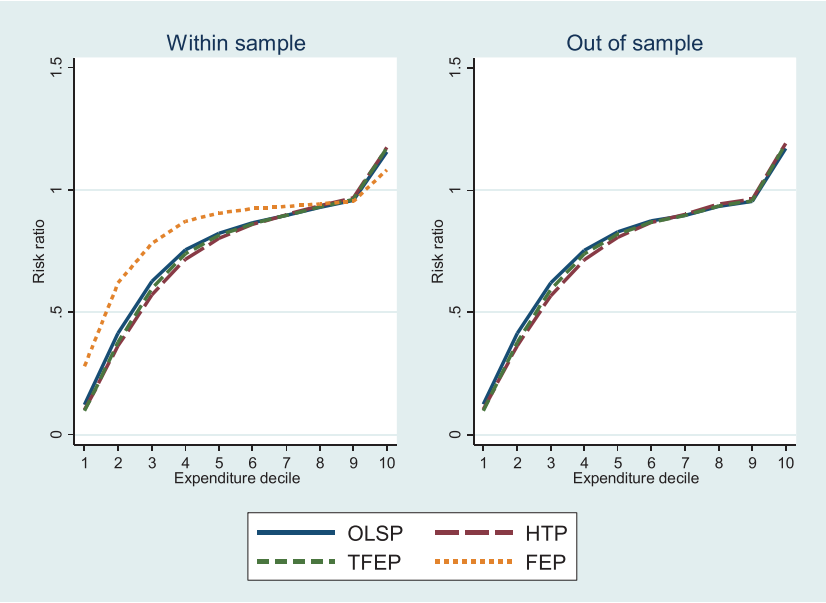


Figure 3. Predictive risk ratios for out-of-sample prediction with concurrent specifications

The performance of the feasible concurrent predictors in Figure 3 is very similar, although the OLSP does better than either TFEP or HTP for most of the deciles. There is little degradation in performance out of sample relative to within sample for these predictors but this simply confirms the representativeness of the chosen holdout sample.

Figure 4 again highlights the poorer performance of the prospective model relative to that of the concurrent specification. Just as in the case of the concurrent specification, the OLSP tends to outperform the available alternatives—TFEP and HTP—for most deciles.

DISCUSSION

Overall, the results of the risk adjustment case study point to a clear preference for the OLSP over all of the variants of predictors based on FE (FEP, TFEP and HTP) that were considered. There were some exceptions where FE-based

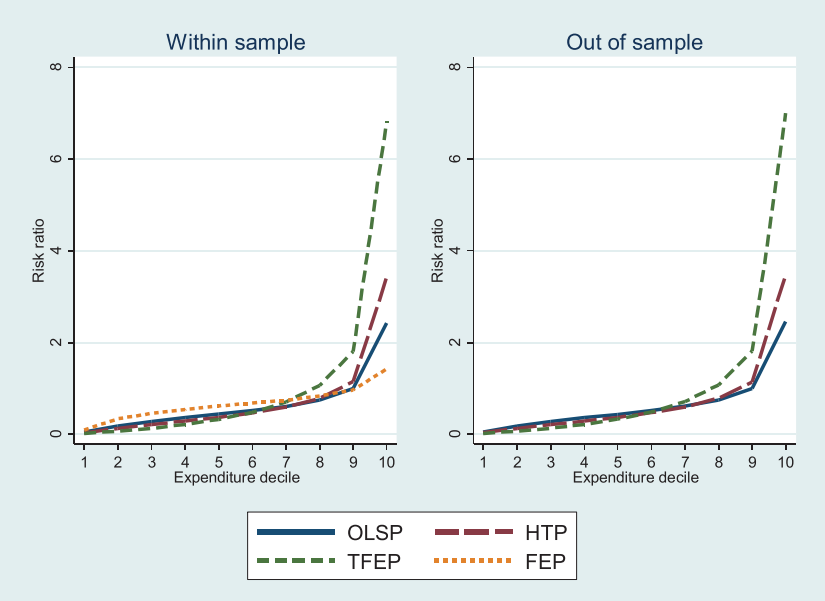


Figure 4. Predictive risk ratios for out-of-sample prediction with prospective specifications

predictors proved superior but these were limited to some of the comparisons made on the basis of risk ratios (group-level prediction).

The superiority of the OLSP for out-of-sample prediction was anticipated from the simulation evidence that was presented. For FE-based predictors to be competitive they need to include estimated individual effects, otherwise the possibly better (less biased) estimates of the coefficients of the time-varying variables are detrimental to producing good forecasts. In contrast, the OLSP will be the best linear predictor conditional on the available regressors. One might anticipate that incorporating time-invariant regressors to allow for estimated individual specific effects might favourably impact on the relative performance of the FE-based predictors. This conjecture was considered in our analysis of health expenditures. Using an extensive list of extra time-invariant regressors to predict individual specific effects, the resultant HTP did indeed provide some improvement over the TFEP for use out of sample, but this improvement was not enough to overcome the superiority of OLSP.

Somewhat surprising, given the simulation evidence, was the relative performance of predictors in the case of post-sample prediction. Using the estimated individual specific effects in time series forecasting did not uniformly improve the forecasting performance of FEP, and the OLSP was the best-performing predictor in all cases when predictive MSE was the performance measure. This result proved sensitive to the performance measure and, when predictive ratios were used to evaluate post-sample performance, the FEP did outperform the OLSP. This was the only situation where this ranking occurred.

The relative performance of the OLSP and FEP in post-sample prediction is impacted by a number of factors. In the simulation evidence, FEP did improve monotonically with ρ, the proportion of the variability in the unobservables attributable to the individual specific effects. Holding constant the signal-to-noise ratio means the OLSP is unaffected, and so for small enough values of ρ the OLSP is expected to dominate FEP. In our analysis of health expenditures, the estimated ρvalues were in the range 0.2–0.5, depending on the sample and whether the prospective or concurrent specification was used. So, while in general this is a possible explanation of the relatively poor performance of FEP, these estimates of ρ suggest this is not the explanation here.

Instead, two other contributing factors closely associated with micro panel data are likely to be the main reasons for the relatively poor performance of FEP. The first of these is the extent of within-variation, which, if limited, as it was in the risk adjustment case study (and likely to be in most micro panel data), will result in deterioration in the relative performance of FE-based predictors through relatively poor estimates of the coefficients of the explanatory variables. The second and possibly most important factor that influences the relative performance of predictors is the quality of the estimates for the individual specific effects. The dominance of FEP within sample derives from the inclusion of the estimated individual effects. But even though non-zero individual effects may exist, in cases such as ours with small T available to estimate these effects, they are likely to be estimated with considerable variability, which contributes to higher predictive forecast variability. The consequence is that a simpler model, here OLSP, is likely to prove superior in terms of predictive MSE.

This result, where the simpler OLSP performs well, is consistent with evidence drawn from the literature comparing homogeneous and heterogeneous panels, i.e. whether coefficients on regressors are allowed to vary over individuals or not. Such comparisons require panels with at least modest T values and so are not strictly transferable to the micro panel case. But there is a common theme that emerges in this literature, which Baltagi (2008, p. 169) describes as follows:

> … although the performance of various panel data estimators and their corresponding forecasts may vary in ranking from one empirical example to another, … the consistent finding in all studies is that homogenous panel data estimators perform well in forecast performance mostly due to their simplicity, their parsimonious representation and the stability of parameter estimates.

This type of result is highlighted in Clark and McCracken (2012), who formalize the general issue of trade-offs in forecast accuracy associated with noise in parameter estimation.

The other dimension of the comparisons made is that between concurrent and prospective specifications. The superiority of a concurrent model in terms of lower forecast errors is known. Using German data, the MAPE from a prospective model is more than 60% greater than that from the concurrent model (Behrend *et al.*, 2007). Using Taiwan data, the corresponding figure is about 25% (Chang and Weiner, 2010). The results presented here demonstrate similarly large differences in performance. This is primarily because the concurrent specification captures more of the costs of actual utilization during a year. However, from a post-sample forecasting perspective such a specification does not represent a truly operational predictive tool. On the other hand, the prospective specification relies on past factors and so predictions of future utilization are readily generated. While the concurrent model more accurately reflects actual spending, for payers using risk adjustment models the prospective model gives advance indication of what their financial obligations will be. Thus, for payment purposes, the prospective model is used more often than the concurrent model; concrete applications include social health insurance in Germany and Medicare Shared Savings program in the USA.

To put the concurrent specification on an equal footing with the prospective specification requires the use of predicted covariates to make the associated predictors operational. This is not an approach implemented here and the question of which of these approaches is better is left for further research. What has been emphasized in this paper is the relative performance of alternative predictors when used in conjunction with different specifications and for use in post-sample and out-of-sample prediction.

CONCLUSION

A rich dataset comprising the linkage of a large cohort-representative survey to several years of comprehensive health records provides a test bed to explore the relative forecasting performance of alternative models of health expenditures. In contrast to much of the risk adjustment literature, where such modelling is prevalent, our focus is on predictors that exploit the availability of panel data. We also stress the distinction between predictions made out of sample and post sample that is possible when panel data are used. While it is unwise to draw strong conclusions on the basis of our single case study and a somewhat limited extension of the simulation results of Baillie and Baltagi (1999), there are a number of general issues that our work highlights in terms of forecasting with micro panels that feature a large number of individuals but limited time periods.

First, the strong preference for FE over OLS estimators when parameter estimation is the goal does not readily extend to situations where forecasting is the main task. We demonstrate that this preference is fragile, and is likely to be overturned in many practical situations with micro panels, including our models of health care costs. Simulation results add support and additional insights into the results obtained in the application. These results are supportive of the use of the OLSP in a wide range of circumstances.

Second, while FE-based predictors may prove useful in other applications, favourable circumstances would need to exist for this to happen. These would include a relatively high proportion of the variability in the unobservables attributable to individual specific effects and having regressors with considerable within-variation. A third factor that would be helpful is to have a relatively large number of time series observations in order to better estimate individual specific effects. However, such a situation takes us out of the realm of micro panels into a situation where alternative approaches and model specifications might be entertained.

Despite the limited success of predictors that explicitly utilize the panel structure, there are advantages of having panel data that have not been highlighted but nonetheless need to be recognized. First, the prospective model that arguably is the more useful specification for risk adjustment requires panel data. Second, there are likely to be gains in predictive performance that derive from simply having extra data. More specifically in the particular case of risk adjustment, the key set of predictors is associated with hospital diagnoses, which come sporadically. Panels provide richer data in this very important domain.

ACKNOWLEDGEMENT(S)

This research uses data from the 45 and Up Study, which is managed by the Sax Institute in collaboration with major partner Cancer Council New South Wales; and partners the Heart Foundation (NSW Division); NSW Ministry of Health; *beyondblue*; Ageing, Disability and Home Care, NSW Family and Community Services; Australian Red Cross Blood Service and UnitingCare Ageing. This project was undertaken by the University of Technology Sydney and utilized Medical Benefit Schedule (MBS) and Pharmaceutical Benefit Schedule (PBS) data supplied by the Department of Human Services. Data linkage for the project was undertaken by the Centre for Health Record Linkage. The 45 and Up Study has the approval of the University of NSW Health Research Ethics Committee; this project has ethics approval from the NSW Population and Health Services Research Ethics Committee and the Department of Health and Ageing Departmental Ethics Committee. The study's findings are those of the authors and do not necessarily represent the views of the Department of Health and Ageing, or the Department of Human Services. The project is funded by an ARC Discovery Project grant (DP110100729). We are grateful for the helpful comments made by Randy Ellis and seminar participants at the universities of New South Wales, Monash, Flinders and Wollongong.

REFERENCES

45 and Up Study Collaborators. 2008. Cohort profile: the 45 and Up Study. *International Journal of Epidemiology* **37**(5): 941–947.

Aizcorbe A, Liebman E, Pack S, Cutler D, Chernew M, Rosen A. 2012. Measuring health care costs of individuals with employer-sponsored health insurance in the U.S.: a comparison of survey and claims data. *Statistical Journal of the IAOS* **28**(1): 43–51.

Albouy V, Davezies L, Debrand T. 2010. Health expenditure models: a comparison using panel data. *Economic Modelling* **27**(4): 791–803.

Ash AS, Zhao Y, Ellis RP, Kramer MS. 2001. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. *Health Services Research* **26**(6): 194–206.

Baillie RT, Baltagi BH. 1999. Prediction from the regression model with one-way error components. In Analysis of Panels and Limited Dependent Variable Models, Hsiao C, Lee LF, Pesaran H (eds). Cambridge University Press: Cambridge, UK; 225–267.

Baltagi BH. 2008. Forecasting with panel data. *Journal of Forecasting* **27**(2): 153–173.

Behrend C, Buchner F, Happich M, Holle R, Reitmeir P, Wasem J. 2007. Risk-adjusted capitation payments: how well do principal inpatient diagnosis-based models work in the German situation? Results from a large data set. *European Journal of Health Economics* **8**: 31–39.

Bernard D, Farr S, Fang Z. 2011. National estimates of out-of-pocket health care expenditure burdens among nonelderly adults with cancer: 2001 to 2008. *Journal of Clinical Oncology* **29**: 2821–2826.

Chang HY, Weiner J. 2010. An in-depth assessment of a diagnosis-based risk adjustment model based on national health insurance claims: the application of the Johns Hopkins Adjusted Clinical Group case-mix system in Taiwan. *BMC Medicine* **8**: 7.

Clark TE, McCracken MW. 2012. In-sample test of predictive ability: a new approach. *Journal of Econometrics* **170**: 1–14.

Einav L, Finkelstein A, Ryan S, Schrimpf P, Cullen M. 2013. Selection on moral hazard in health insurance. *American Economic Review* **103**(1): 178–219.

Ellis RP. 2008. Risk adjustment in health care markets: concepts and applications. In Financing Health Care: New Ideas for a Changing Society, Lu M, Johnson E (eds). Wiley: Chichester; 177–222.

Ellis RP, Fiebig DG, Johar M, Jones G, Savage E. 2013. Explaining health care expenditure: large-sample evidence using linked survey and health administrative data. *Health Economics* **22**: 1093–1110.

Hausman JA, Taylor WE. 1981. Panel data and unobservable individual effects. *Econometrica* **49**(6): 1377–1398.

Hill SC, Miller GE. 2011. Health expenditure estimation and functional form: applications of the generalized gamma and extended estimating equations models. *Health Economics* **19**: 608–627.

Hsiao C. 1986. Analysis of Panel Data. Cambridge University Press: Cambridge, UK.

Jones A. 2009. Panel data methods and applications to health economics. In *Palgrave Handbook of Econometrics*, Vol. 2, Mills TC, Patterson K (eds). Palgrave MacMillan: London; 557–631.

Jones A. 2011. Models for health care. In Oxford Handbook of Economic Forecasting, Hendry D, Clements M (eds). Oxford University Press: Oxford; 625–654.

Jones A, Lomas J, Rice N. 2014. Applying Beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics* **29**(4): 649–670.

NSW Health. 2011. Cost of Care Standards 2009/10. Sydney: NSW Health.

Seshamani M, Gray J. 2004a. A longitudinal study of the effects of age and time to death in hospital costs. *Journal of Health Economics* **23**(2): 217–235.

Seshamani M, Gray J. 2004b. Ageing and health-care expenditure: the red herring argument revisited. *Health Economics* **13**(4): 303–314.

Stearns S, Norton E. 2004. Time to include time to death? The future of health care expenditure predictions. *Health Economics* **13**(4): 315–327.

Wooldridge JM. 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.

Zhao Y, Ash AS, Ellis RP, Ayanian JZ, Pope GC, Bowen B, Weyuker L. 2005. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Medical Care* **43**(1): 34–43.

Zuvekas S, Olin G. 2009. Accuracy of Medicare expenditures in the Medical Expenditure Panel Survey. *Inquiry* **46**(1): 92–108.

APPENDIX

HEALTH EXPENDITURE DATA AND REGRESSIONS

Table A1. Summary statistics of selected variables

Variable	Mean	Variable	Mean
This year's expenditure	4.150	**Ever diagnosed**	
(SD)	9.212	High BP	0.356
Next year's expenditure	4.420	Skin cancer	0.285
(SD)	9.807	Breast/prostate cancer	0.058
Age	62.213	Other cancer	0.062
(SD)	11.150	Heart disease	0.118
Male	0.463	Stroke	0.031
Never married	0.062	Diabetes	0.090
Widowed	0.085	Asthma/fever	0.024
Separated	0.073	Depression	0.130
Married*	0.692	Broken bone	0.115
Divorced	0.028	Urinary leakage	0.351
Unknown	0.006	**Hospitalized conditions**	
Partner	0.054	Infections	0.014
Foreign language	0.096	Solid tumors	0.014
Foreign born	0.251	Diabetes II	0.023
Health card	0.295	Eye conditions	0.026
Education: high school*	0.134	Benign/uncertain neoplasms	0.028
Education: certificate	0.318	Bladder and urinary conditions	0.023
Education: diploma	0.318	Cardiac arrhythmias	0.015
Education: university	0.230	Coronary artery disease	0.017
Major city*	0.450	Gastrointestinal conditions	0.074
Remote	0.020	Hyperlipidemia and Lipidoses	0.012
Outer region	0.178	Hypertension	0.037
Inner region	0.352	Lung diseases	0.012
SAH: excellent*	0.146	Musculoskeletal condition	0.045
SAH: very good	0.357	Skin condition	0.013
SAH: good	0.326	**Prescriptions**	
SAH: fair	0.115	Angiotensin converting enzyme inhibitors	0.113
SAH: poor	0.020	Angiotensin II inhibitors	0.092
PHI with extra	0.491	Antianginal agents	0.032
PHI no extra	0.145	Anticoagulants (warfarin)	0.029
No PHI*	0.365	Antidepressants (non-SSRI)	0.060
LF: other work	0.147	Antidepressants (SSRI)	0.052
LF: full-time*	0.339	Antigout agents	0.037
LF: fully retire	0.374	Antihypertensive combinations	0.084
LF: disabled	0.042	Anti-infectives (oral)	0.259
LF: not working	0.099	Antiplatelet agents	0.078
BMI: underweight	0.013	Asthma, COPD (inhaled beta-agonist)	0.097
BMI: normal*	0.339	Beta-adrenergic blocking agents	0.093
BMI: overweight	0.366	Calcium channel blocking agents	0.107
BMI: obese I	0.148	Lipid-lowering agents (statin)	0.271
BMI: obese II	0.042	Non-steroidal anti-inflammatory agents	0.144
BMI: obese III	0.017	Oral diabetic agents	0.052
Current smoker	0.072	Topical steroids/anti-inflammatories	0.094
Past smoker	0.351	Ulcer/GERD (PPI)	0.212
Non-smoker	0.576	Osteoporosis treatments	0.054

Note: *Reference category. Age enters the regression model in 5-year band categories. Also included in the regression are year dummies, interaction terms between age categories and sex, dummies for skin colour, income categories, alcohol habits, other hospitalized condition dummies (110 in total) and other prescription drug dummies (121 in total).

Authors' biographies:

Denzil G Fiebig is a Professor in the School of Economics at the University of New South Wales in Sydney Australia. His research interests include micro-econometrics, forecasting and health economics.

Meliyanni Johar is a research Associate Professor at the Economics Discipline Group at the University of Technology Sydney, Australia. Her research is mainly in applied econometrics with a focus in health economics. She has published in top journals such as the Journal of Health Economics, the Journal of Applied Econometrics, the Social Science and Medicine, and Health Economics.

Authors' addresses:

Denzil G. Fiebig, School of Economics, University of New South Wales, Sydney, Australia.

Meliyanni Johar, Economics Discipline Group, University of Technology Sydney, Australia.

