Centre for Health Economics and Policy Analysis

CHEPA WORKING PAPER SERIES

Paper 15-05

Forecasting Health Expenditure: Methods and Applications to International Databases

2015

Junying Zhao[1]

[1] Health Policy PhD Program, McMaster University, Hamilton, ON, Canada. zhaoj6@mcmaster.ca.

Acknowledgements: I am indebted to Jeremiah Hurley and Thomas Getzen for detailed and constructive comments. I thank Arthur Sweetman for remarks and Michel Grignon for recommending a data-analysis opportunity of health accounts for the World Bank. I acknowledge the learning experience about National Health Accounts at the World Health Organization.

McMaster University

CHEPA WORKING PAPER SERIES

The Centre for Health Economics and policy Analysis (CHEPA) Working Paper Series provides for the circulation on a pre-publication basis of research conducted by CHEPA faculty, staff and students. The Working Paper Series is intended to stimulate discussion on analytical, methodological, quantitative, and policy issues in health economics and health policy analysis. The views expressed in the papers are the views of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of Working Papers are encouraged to contact the author(s) with comments, criticisms, and suggestions.

NOT FOR CITATION WITHOUT PERMISSION

Abstract

This paper examines a number of issues encountered when using standard health accounts data to forecast national health expenditures. In particular, it focuses on measurement issues, model specifications, and a comparison of performance indicators based on commonly used health accounts data from OECD. It assesses the performance of alternative forecasting methods based on three criteria — accuracy, precision, and certainty. Based on these criteria, it assesses the performance of model specifications including univariate (i.e., health spending) and multivariate (e.g., macroeconomic factors), static (e.g., fixed effect) and dynamic (e.g., dynamic panel), and single-equation models (e.g., ARIMA) and system of equations (e.g., VAR). It uses the better-performing models to forecast health expenditures for individual countries. This analysis makes three contributions to the literature on forecasting health expenditures. First, with longer data series, in contrast to some previous papers on health expenditure projections, we obtain a result that is more conventional in the forecasting field — econometric time series models and statistical smoothing models perform better than econometric panel data models. Second, a recent literature review of health expenditure forecasting suggests that with better computing power and more refined data, the future of forecasting is complicated micro models. But modeling and understanding the determinants of expenditure growth (whether using micro data or CGE macro models) require considerably more data and effort, and may still do worse with pure forecasting. This paper confirms this. At the same time, it contributes to the call for more rigorous methods of forecasting, for more transparency, and for better assessment of performance. This analysis can inform both research and policy debate on budgetary planning and fiscal sustainability of health expenditure.

JEL classification: C53; E17; H51; I1

Keywords: Forecasting; Accuracy; Precision; Health Expenditure; System of Health Accounts

1 Introduction

Health expenditures have been increasing in the past decades among developed and developing countries (Newhouse, 1977; Culyer, 1988; Gerdtham *et al.*, 1992; Panopoulou and Pantelidis, 2011; Lago-Peñas *et al.*, 2013; Lorenzoni *et al.*, 2014). This raises global concerns about health and the economy, and calls for fiscal preparations and policy interventions to ensure future fiscal sustainability (European Union, 2009; Centers for Medicare & Medicaid Services, 2011). Forecasting health expenditures is crucial for policy applications required by governmental organizations and central banks. For example, under-predicting public health expenditures can result in unmet health needs that are eligible for public support, or in shortages of infrastructure investments in hospitals and health human resources. Excess public health expenditure on hospitals and physician services crowds out financial resources that might have been allocated to other equally important sectors such as education for human capital accumulation. It may also absorb labor that could have been directed to the final goods sector and hence reduce manufacturing outputs, which in turn may hurt both welfare and economic growth over time.

This paper primarily addresses the methodological question: how can we best forecast health expenditure in a systematic way? It then answers the empirical question: given our best estimates, how much would worldwide health expenditures be in the short- and medium-run future (e.g., 2015-2025)? This paper examines a comprehensive set of measurements and projection models for forecasting health expenditures, and carefully tests their forecast performance based on formal criteria and by using recent and comparable data provided by international sources. The methods and empirical results of health expenditure forecasts can inform the policy making process by projecting needed funds and identifying gaps between the amounts of monetary resources needed and those available.

Literature relevant to these questions involves two sub-fields of economics: health economics (esp., expenditure), and economic forecasting. The majority of the health expenditure literature tries only to understand past drivers of health expenditure, rather than project health spending into the future. Although we are aware that there are hundreds of such past-estimation articles using micro data, we only focus on those using aggregate data since our data is at the country level. The seminal paper by Newhouse (1977) regressed per capita health expenditure on per capita GDP using cross-sectional data from OECD countries, and found that GDP is a statistically and economically significant factor in explaining health spending. Empirical studies using the OECD database have tried various econometric models and estimation methods. Gerdtham and Jonsson (2000) divided studies into two generations. The first-generation studies used cross-sectional data for a single year (or selected years) to examine within-country determinants of, and cross-country differences in, health expenditure. While these studies paid attention to methodological issues such as the appropriate choice of the currency unit (or conversion factors) (Parkin *et al.*, 1987, 1989), they have been criticized owing to shortcomings inherent in cross-sectional data estimates. For example, small sample size and omission of the country- and time-invariant variables can generate inconsistent estimates of the regression coefficients. These criticisms were addressed by the second-generation studies that used panels of countries, where a relatively long time series of annual data was available for each country. These latter studies, however, faced other methodological issues such as non-stationarity[1], cointegration[2] between two and more variables, and heterogeneity across countries (Gerdtham *et al.*, 1992; Barros, 1998).

[1]Non-stationarity is defined in the literature in various but similar ways: "Data in economics is often nonstationary, namely, has changing means and variances over time. Failure to allow for such specific characteristics may result in inferior forecasts of aspects of interest " (Clements and Hendry, 2004) (p.8). "Stationarity means that ... the data fluctuate around a constant mean, independent of time, and the variance of the fluctuation remains essentially constant over time " (Makridakis *et al.*, 1998) (p.324).

[2]Cointegration intuitively means that a long-run relationship exists between variables.

In addition, the broader economic literature on forecasting emphasizes fundamental methodological challenges as well. First, forecasting imposes an assumption of continuity (NBER, 1966; Theil, 1966) — that the pattern of data in the past persists in the future. Second, an explanatory forecast ignores the unquantifiability of certain variables such as technological progress. Third, forecasting has to acknowledge the "unknown uncertainty" (Clements and Hendry, 2004, 2011) that we have not discovered and hence are unable to incorporate the true data generating process. Rigorous forecasts, however, can be achieved from sound economic theories and econometric methods, and careful assessment of the performance of empirical models according to three formal criteria — accuracy, precision, and known uncertainty (Clements and Hendry, 2004, 2011). Forecasts based on sound grounds do add new information about the status of the future, hence can stimulate governmental reactions within a health system (Box *et al.*, 2008). If a forecast depicts a gloomy picture ahead, then policy makers can respond by changing certain conditions to prevent its realization. Rigorous forecasts, better than *ad hoc* guesses, can successfully guide policy makers to enhance the likelihood of a favorable outcome and avoid an undesirable outcome.

Few studies were found to integrate health expenditure analysis and economic forecasting by searching truncated terms "health expenditure AND (project OR forecast)" in two commonly used economic literature databases (EconLit, IDEAS). To forecast provincial health expenditure, Di Matteo (2010) used measures of real per capita Canadian provincial governmental spending on health care from 1965 to 2008, economic and demographic regressors such as provincial GDP, provincial population in total and the proportion of the elderly, federal cash transfers, and provincial governmental revenues and expenditures. The study used classical regression estimators including ordinary (OLS) and generalized least squares (GLS), focused only on regressors with statistically significant coefficient estimates, and used simple extrapolations of their respective historical growth rates into the future to generate future values

of health expenditure. The study neither included specification tests for the consistency of the pooled OLS and GLS estimators (e.g., whether the constant coefficients α and β are appropriate for the purpose of using pooled OLS estimator) nor formally reported forecast performance of the models.

Getzen and Poullier (1992) conducted forecasts of national health expenditure using data for 19 OECD countries. They used estimates based on 1965–1979 data to forecast within the sample (1980-1987) using the naïve method, exponential smoothing, and multivariate (inflation and GDP) models respectively for each country and the pooled-country panel. They then measured the mean absolute error (MAE), an indicator of forecast accuracy, and compared MAE of each method with that of the naïve method. For each country, the combination or average of forecasts obtained from single forecasting models was found to be more accurate than any single model alone. However, the finding that multivariate regressions offered more accurate forecasts than time series models such as ARIMA is inconsistent with much of the literature on economic forecasting (Armstrong, 2001; Makridakis *et al.*, 1998). Such a finding may arise because, as has been recognized, an insufficient number of observations leads to unstable estimated parameters. Overcoming this limitation requires waiting for more data to accumulate. This is the case in the present paper wherein an additional 20 years of data are used to forecast.

Moreover, Getzen (2006, 2007) has argued that model specifications of health expenditures vary with temporal and spatial dimensions, that is, the chosen length of time horizon and breadth of the observational unit. He presents a framework to categorize health expenditure projection methods using conceptual metrics that consists of four aspects: observational units at the macro- and micro-levels; time span over the short, medium, and long run; measures of Total Health Expenditure (THE); and incorporation of various regressors. The best indicator of the short-run growth of nominal THE was its growth rate in the previous year, plus a time

trend, and a few regressors including employment and inflation rates, rather than other factors. Endogenous and pre-determined variables such as health, demographics (e.g., aging), hospital infrastructure, and physician supply change little in the short run (i.e., 1–3 years), and other factors have already been constrained within the existing budget (e.g., funds for research and development, affordability of technology adoption). Therefore, their effects on THE neutralize overall, and their attributions to the short-term forecast of THE are argued to be negligible.

Good indicators of medium-run (i.e., 3 –10 years) forecasts of THE can be achieved using only past values of THE, plus one and only one crucial regressor — national income — both measured in real, per capita terms. This approach was adopted for this analysis for medium-run forecasts. We are not interested in the long run since anything can occur over such a long period.

As Astolfi *et al.* (2012) point out, at the macro (e.g., country) level, there is another branch of forecasting that uses structural models — computable general equilibrium (CGE) models. These models rely on economic theory, which is helpful in selecting variables as potential drivers of HE and in imposing assumptions (e.g., existence and number of equilibria in CGE models) that may decrease parameters' estimation error (Elliott and Timmermann, 2008). This type of model is particularly helpful to answer "what if" questions regarding the results of exogenous policy interventions. In contrast, because they do not specify mechanisms for policy and changes in both health and non-health sectors, forecasts based on reduced-form models cannot clearly explain "why" and "how" an effect occurs. However, the CGE approach is both more difficult and more costly because it requires the construction of formal structural models of health and the economy. Constructing such a structural model is beyond the scope of this paper.

Besides forecasts from academia, governmental agencies also provided forecasts. For instance, the U.S. Medicare Trustees provide detailed short-term projections of

health sub-sectors, and long-run forecasts of total Medicare costs. They first project the percentage growth rate of real per capita GDP, which is then used to project aging-related Medicare costs (Centers for Medicare & Medicaid Services, 1991). They then add 1% to incorporate a technology effect (Centers for Medicare & Medicaid Services, 2000, 2004). This has been criticized as arbitrarily adding an excess growth rate of GDP (Getzen, 2007). Without being tested, aging was deemed by the OECD (2003) and the European Union (2008) as a major determinant of health care costs and an attribute to forecasts of health expenditures. It has been argued that aging *per se* is neither a speedy nor a substantial driver of health care cost growth, compared with the effects of growth of the national income and budget (Denton *et al.*, 2002; Evans *et al.*, 2001). Therefore, it is useful to forecast health expenditures both excluding (e.g., in univariate time series models) and including (e.g., in panel data models) demographic variables separately.

To sum up, Getzen and Poullier (1992) obtained an unusual result — that panel data models perform better than time-series models — and argued that people should examine this issue as more data accumulate. Our paper does exactly this and obtains the more conventional result, in a sense, overturning Getzen and Poullier's initial finding. The comprehensive literature review conducted by Astolfi *et al.* (2012) suggested that with better computing power and more refined data, the future of forecasting is complicated micro-level models. The objective of simply getting accurate forecasts of future spending from modeling, however, needs to be carefully distinguished from that of understanding the determinants of expenditure growth. Modeling and understanding the determinants of expenditure growth (whether using micro data or CGE macro models) require considerably more data and effort, and may still do worse for pure forecasting. Our paper confirms this. At the same time, our paper contributes to the call (articulated in Astolfi *et al.* (2012)) for more rigorous methods, for more transparency, and for better assessment of forecast performance.

2 Specification of Projection Models

Projection models require three basic assumptions: past information is available, it is quantifiable, and certain characteristics of the pattern of such data continue into the future, i.e., the assumption of continuity or constancy (NBER, 1966; Theil, 1966). A short-term forecast usually spans 1 to 3 years, whereas intermediate and long-range predictions are typically 3 to 10 years and beyond 10 years, respectively (Getzen, 2000). Generally, the projection model in simple logarithmic form is

$$lnY_t^f = X_t b \quad (\text{t=T+1,T+2,}\ldots\text{,T+F}), \tag{1}$$

where T means the time horizon of the estimated past data; F refers to the number of forecast periods; Y is the dependent variable that is to be forecasted at a point in time as Y_t^f; and X is the independent variable that is used to forecast Y. $\mathrm{E}(ln(Y_t^f)) = X_t\mathrm{b}$. $\mathrm{Var}(ln(Y_t^f)) = X_t[\mathrm{Var(b)}]X_t'$. The forecast error $e_t = ln(Y_t)$-$ln(Y_t^f)$ (t=T+1,T+2,...). $\mathrm{E}(e_t) = 0$ and $\mathrm{Var}(e_t) = s_t^2$ where s is the estimated standard error. Given a level of significance $\alpha > 0$, the confidence interval of forecast of $ln(Y_t)$ is defined using

$$\Pr\left[lnY_t^f - t_{\frac{\alpha}{2}} s_t \leq lnY_t \leq lnY_t^f + t_{\frac{\alpha}{2}} s_t\right] = 1 - \alpha.$$

Based on the taxonomy of projection models presented by Clements and Hendry (2004), we categorize projection methods according to the following aspects: time span (short vs. medium vs. long run); data pooling (single- vs. pooled-country); type of model (statistical smoothing vs. econometric time series data vs. panel data); number of independent variables (univariate vs. multivariate); existence of causality (non-causal vs. causal[3]); existence and number of time lags (static vs. dynamic);

[3]Admittedly theorists who favor structural models may be less likely to agree with empiricists′ causal interpretation of certain reduced-form models.

and directness of forecast procedure (direct vs. combined). Based on meaningful combinations of these aspects, we divide six commonly used reduced-form models into three time series and three panel data models (Table 1). Each model features certain of the above-listed categories. As we consider the six projection models below, each is found to overcome certain methodological weaknesses of the preceding model.

Table 1: Projection models

Model class	(Estimation) Model specifications
Time series data	
M1. Exponential smoothing	$y_t^f = y_{t-1}^f + \alpha(y_{t-1} - y_{t-1}^f)$
M2. ARIMA(p,I,q)	$y_t = b_0 + b_1 y_{t-1} + \ldots + b_p y_{t-p} + e_t - d_1 e_{t-1} - \ldots - d_q e_{t-q}$
M3. VAR	$Y_t = A_0 + A_1 Y_{t-1} + \ldots + A_p Y_{t-p} + e_t$
Panel data	
M4. Static, fix effects	$y_{i,t} = a_{i,fe} + \gamma_1 x_{1i,t} + \ldots + \gamma_n x_{ni,t} + e_{i,t}$
M5. Dynamic	$\Delta y_{i,t} = a + b_1 \Delta y_{i,t-1} + \ldots + b_p \Delta y_{i,t-p} + \gamma_1 x_{1i,t} + \ldots + \gamma_n x_{ni,t} + \Delta e_{i,t}$
M6. Nonstationary & cointegrated	$y_{i,t} = \gamma x_{i,t} + e_{i,t}$, where $x_{i,t} = x_{i,t-1} + u_{i,t}$

Notes: f means forecast. α is the weight on previous forecast error. Y refers to the matrix of dependent variable y and independent variable x. A is the matrix of coefficients. Δ represents first difference. $y_{i,t}$ in M6 is cointegrated with $x_{i,t}$, and $x_{i,t}$ is integrated process of order 1 for all i. Letters e and u refer to different error terms. Parameter p means time lag of a variable. q is the time lag of error. γ is the coefficient of a variable. Estimated values of the same parameter are not necessarily the same in differently specified models.

The exponential smoothing model (M1) requires only a single series. Therefore, it produces forecasts with little information at low costs. It aims to reduce randomness and estimate the trend component of a series. Observations are given weights that exponentially decline as observations become remote (Holden *et al.*, 1990). Both double exponential (DE) smoothing and its alternative Holt-Winters non-seasonal smoothing are conducted to project the series mean (and hence the point forecast) rather than the variance (and hence the forecast interval), since the latter becomes invalid when the assumption of independently-and-identically-distributed errors (i.i.d.) does not hold. The autoregressive integrated moving average model ($AR_pI_dMA_q$, M2) is argued by Granger (1989) to improve forecasts with low marginal cost since the information it requires is still about a single variable and with high marginal benefit as one more recent (rather than average) observation gives better information for future forecasts. To confirm estimates of the number of Y′s own lagged values, p,

and the number of current and lagged residuals, q, we use Durbin-Watson, Durbin′s alternative, and Breusch-Godfrey tests for lower- and higher-order serial correlations of residuals. To diagnose d, the number of times of differencing, it is necessary to difference the series into stationarity. We use the Augmented Dickey-Fuller (ADF) test and other two tests for unit roots — the Philips-Perron (Philips and Perron, 1988) and DF-GLS (Elliott *et al.*, 1996) tests. ADF imposes the strict assumption that errors are serially uncorrelated and homogeneously distributed, while Phillips-Perron allows errors to be weakly dependent and heterogeneously distributed. The DF-GLS test successfully distinguishes effects of unit roots from the deterministic trend and hence has greater power than ADF.

The vector autoregression (VAR, M3) model allows for endogeneity and bi-directional correlation (not necessarily causality (Sims, 1972)). The cointegration between variables (e.g., THE and GDP) signifies further possible cointegration between subsets of these variables (Dolado *et al.*, 1990), which justifies our examination of the relationship between public health expenditure (PHE) and general governmental revenue (GGR). Multicollinearity is not a concern as long as the assumption of continuity (i.e., the pattern of such multicollinearity continues into the future) holds. All series are checked for cointegration by applying ADF, Philips-Perron, and DF-GLS tests to the residuals. The stability of VAR and the normality of disturbances distribution are also tested. The model that passes all specification tests is used for estimation, point forecast, and forecast interval. If more than one adequate specification occurs, we select the better-estimated one according to the principle of parsimony using minimum information criteria (Akaike′s, AIC; Bayesian, BIC) for lags and the three criteria of forecast performance (introduced in the next section). We do not estimate the Vector-Error-Components (VEC) model because for some countries the GDP (or GGR) is not cointegrated with THE (or PHE) through the Johansen's cointegration test. This justifies using VAR generally and consistently for all countries rather than

VEC conditional on cointegration.

Compared to time-series data models, panel data offers three advantages: a larger number of observations which enhances estimation precision, the possibility of establishing causation under the assumptions of the fixed effects (FE) model (Cameron and Trivedi, 2010), and the ability to estimate country-specific effects. However, both the number of countries N and the number of years T of the series at the annual national level are relatively small, which does not allow formal tests for poolability and country-specific effects (Cameron and Trivedi, 2010). We specify a constant slope β, as is commonly done in the literature when using such aggregate data, and use the Hausman test to indirectly test the consistency of Random Effects (RE) by comparing it with FE. Other estimators are disregarded because the "within" estimator is equivalent to FE if the intercepts are fixed effects and errors are i.i.d., while "pooled" and "between" estimates are similarly consistent with and usually less efficient than those obtained from RE (Cameron and Trivedi, 2010). We find that RE is rejected by the Hausman test and use panel-robust standard errors that correct both serial correlation and heteroskedasticity.

Such static models, however, impose a strong assumption of exogeneity[4] that can be relaxed by dynamic models. The latter further extract the state dependence of a country — as distinct from unobservable heterogeneity across countries — from the observed total variations of a country. For instance, let y and x be THE and GDP respectively. If FE were true, then true state dependence would attribute the currently high level of THE in a given country to its own high historical level, after controlling for regressor GDP, rather than to the significant difference of this country from others. Therefore, dynamic models can explain state dependence that static models cannot. If state dependence were true for a country, then dynamic models would suggest a different corresponding policy that focuses on the historical problem of this country

[4]The errors have mean zero conditional on the past, current, and future values of regressors.

only. We apply estimators proposed by Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998). These country-panel models, similar to time-series data models, confront the nonstationarity problem since they incorporate the time series component in pooled cross-sections. We diagnose such nonstationarity and cointegration, and estimate using cointegrated methods that have been explored in the general econometrics literature (Baltagi, 2008). In particular, we choose the method proposed by Levin *et al.* (2002), which considers unit root tests in small panels[5] like ours. If nonstationarity is present, we then use the Westerlund (2007) test for the null of no cointegration. If this null is rejected, we then estimate the cointegrated panel after choosing between pooled mean-group (PMG) and mean-group (MG) estimators using the Hausman test. We do find stationarity (consistent with earlier studies (Narayan and Popp, 2012)) so that there is no need to test cointegration.

3 Performance Criteria for Economic Forecasting

A fundamental challenge is how to measure the performance of a particular projection model into the future, rather than simply the "fit" of an estimation model into the past data. Three formal criteria are well established in the field of economic forecasting — accuracy, precision, and certainty (Clements and Hendry, 2004, 2011) — among which, the first is usually treated as the overriding criterion. Accuracy refers to the unbiasedness of the forecast Y^f[6]. Assuming that the intercept equals 0 for simplicity, accuracy is determined by the unbiasedness of estimated parameters: the estimated coefficient of X (i.e., b). Precision means the small size of the variance (σ^2) of the

[5] Other methods such as Im *et al.* (2003); Harris and Tzavalis (1999) require both (infinitely) large N and T for the test statistic to have a well-defined asymptotic distribution.

[6] "The notion of unbiasedness, whereby forecasts are centered on outcomes, is used in technical analyses to measure accuracy; whereas that of small variance, so only a narrow range of outcomes is compatible with the forecast statement, measure precision ... When (squared) bias and variance are combined one-for-one, we obtain the commonly-reported mean square forecast error (MSE)." (Clements and Hendry, 2004) (p.5)

forecast Y^f. Accuracy and precision are often combined into one indicator. In general, there are three alternative groups of such indicators: mean error (ME), mean absolute error (MAE), and mean square error (MSE) (Makridakis *et al.*, 1998)[7]; that is,

$$ME = \sum_{t=T+1}^{T+F} \frac{(Y_t - Y_t^f)}{F},$$

$$MAE = \sum_{t=T+1}^{T+F} \frac{|Y_t - Y_t^f|}{F},$$

$$MSE = \sum_{t=T+1}^{T+F} \frac{(Y_t - Y_t^f)^2}{F}. \tag{2}$$

The sizes of these indicators, unfortunately, depend on the scale of the data. For example, if one used them to select better-performing models where the data are mixed with unlogged and logged scales, then the logged $(Y_t - Y_t^f)$ would be no doubt much smaller than the unlogged $(Y_t - Y_t^f)$ in their numerical values, and one would always mistakenly choose the model using logged data regardless of the model specifications. Because of this, this group of indicators cannot facilitate comparisons across models with different scales of expenditure, which is one of our interests in this paper.

An alternative group of indicators can measure the relative size of forecast error, and hence enable comparisons among various scales of data: percentage error (PE), mean percentage error (MPE), mean absolute percentage error (MAPE), and mean squared percentage error (MSPE); that is,

$$PE = 100 \times \frac{(Y_t - Y_t^f)}{Y_t},$$

[7]F refers to the number of forecast periods.

$$MPE = \sum_{t=T+1}^{T+F} \frac{(PE)}{F},$$

$$MAPE = \sum_{t=T+1}^{T+F} \frac{|PE|}{F},$$

$$MSPE = \sum_{t=T+1}^{T+F} \frac{(PE)^2}{F}. \tag{3}$$

This group of indicators is further subject to two weaknesses. First, if the scale of data includes values close or equal to zero, then the denominator in PE is close to zero, which leads PE to approach infinity. The second disadvantage — shared by both groups of indicators above — is that they represent pure numbers that, without a reference (e.g., naïve model), cannot indicate relative gains in performance using a particular model. For example, ME just gives a mean number of $(Y_t - Y_t^f)$ and MPE offers that of $\frac{(Y_t - Y_t^f)}{Y_t}$. Both measure the distance the forecast Y_t^f of a formal model from the actual reality Y_t rather than from that of another formal model for the very purpose of selecting the better-performing model. The conventional way to proceed is to separately calculate the corresponding MAE, as done by Getzen and Poullier (1992), of the naïve model which uses the most recent observation available Y_t as a one-step forecast Y_{t+1}^f, then compare the indicator based on the naïve model against a formal model, and then select the better performing one. When candidate models are many, this procedure becomes lengthy and inconvenient. Theil′s U statistic, in contrast, is a formal and convenient indicator that builds in the reference naïve method (Theil, 1966; Makridakis *et al.*, 1998) and hence can directly compare the relative performance of candidate models all at once. It also has other favorable characteristics — it penalizes large errors and acts as a compromise between the absolute and relative measures; that is,

$$U = \sqrt{\frac{\sum_{t=T+1}^{T+F}(FPE_{t+1} - APE_{t+1})^2}{\sum_{t=T+1}^{T+F}(APE_{t+1})^2}}, \tag{4}$$

where Forecast Percentage Error: $FPE_{t+1} = \frac{Y_{t+1}^f - Y_t}{Y_t}$ and Actual Percentage Error: $APE_{t+1} = \frac{Y_{t+1} - Y_t}{Y_t}$. Hence the Theil′s U formula can be simplified as

$$U = \sqrt{\frac{\sum_{t=T+1}^{T+F}\left(\frac{Y_{t+1}^f - Y_{t+1}}{Y_t}\right)^2}{\sum_{t=T+1}^{T+F}\left(\frac{Y_t - Y_{t+1}}{Y_t}\right)^2}}. \tag{5}$$

Similar to PE, the FPE and APE[8] respectively measure the predicted relative change of a formal model and the actual relative change of the naïve method. Resembling MSE, both FPE and APE are squared and hence more sensitive to large errors. Finally, the reference of the naïve method is successfully built into APE acting as the denominator of U. Therefore, Theil′s U statistic integrates the methodological advantages of the two alternative groups of indicators, though it does not discriminate between signs like MSE (Theil, 1966). The range and magnitude of U values provide the judgment criteria for forecast accuracy and precision of a model, relative to the naïve reference. If U $>$ 1, the naïve method is preferred to a formal forecast model; if U = 1, both methods are equally good; if U $<$ 1, a formal model is preferred over the naïve method. The closer the U-value is to zero, the more accurate and precise the formal model is. Some may argue that this statistic posits a very simple reference so that it is not hard for a model to perform better than the reference. But this argument is not a concern in our context since what we are comparing are the U statistics between two or more forecast models given the common reference.

[8] Actual refers to the naïve method which uses the actual past value, Y_t, directly for 1-step forecast value, Y_{t+1}^f.

Finally, there is an alternative measure of accuracy and precision that was developed by Diebold and Kilian (2001). They noticed that some variables could be intrinsically more difficult to predict than others. A large gap between forecasted and actual values of such a variable is mainly because the future of the series depends less on its past, rather than the forecaster's failure. It proposes the alternative measure of accuracy and precision as a ratio of the error of a short-run forecast to that of a long-run forecast. As the authors illustrate, this measure is very similar to that of Theil′s U statistic. The main difference is the time horizon. Theil′s U evaluates one-step forecast accuracy of a forecast model relative to that of the naïve model, whereas this alternative measure assesses the one-step forecast accuracy of a forecast model relative to its $k-$step accuracy. So the major strength of this alternative measure is that it allows for different forecast horizons. However, the health expenditures THE and PHE are much less volatile variables, and the international data are almost yearly. That is, HEs are not less-intrinsically-predictable variables and hence are not suitable for applying this alternative measure of accuracy. Therefore, we use Theil′s U statistic as the key performance criterion throughout this paper.

Uncertainty refers to the randomness of the actual Y in the future, which may (i.e., known/measurable uncertainty) or may not (i.e., unknown/immeasurable uncertainty) be captured by the confidence interval of the forecast Y (Clements and Hendry, 2004, 2011). Unknown uncertainty, as noted in the introduction, is due to the limitation of human cognition, and hence cannot be incorporated into models of estimation and prediction. For example, the probability of the Earth being attacked by aliens is beyond human knowledge. Note that this unknown uncertainty is not pertinent to the problem of intrinsic difficulty to predict as described in the preceding paragraph, which refers to the wider confidence interval of volatile variables and is the known uncertainty. The known uncertainty is often measured as the confidence

interval (i.e., the range of certainty)[9] conditional on type I error (α) and on the i.i.d. assumption; that is,

$$(Y^f - z_{\alpha/2}\sigma, \quad Y^f + z_{\alpha/2}\sigma), \tag{6}$$

where σ is replaced by MSE that is obtained from the sample data. It means that the actual Y in the future would have $(100\% - \alpha)$ probability falling within this forecast interval. It remains valid particularly for more than 1-step forecast only when the i.i.d. assumption holds.

4 Strategy of Analysis

We employ an analytical strategy consisting of five stages. In Stage 1 we diagnose the properties of each single time series of THE and PHE containing more than 30 observations between 1960 to 2008 for OECD member countries. Each series is separately measured at two levels — per capita and total; and in three currency units — current national currency units (NCU, nominal), NCU adjusted by GDP deflator in 2000 (NCU, real), and, for the purpose of international comparison, adjusted by Purchasing Power Parity (International Dollars, IntD, PPP). If the series was non-stationary in mean or variance, we transformed it into stationary by differencing or taking the natural logarithm.

In Stage 2, following the rule of thumb in the field of economic forecasting, we estimate parameters in proposed models from the first two-thirds (e.g., 1960-1993) of the historical data (called initialization set). We then use these estimates to predict the dependent variable in the later one-third of data (called test set, e.g.,1994-2008).

[9] Alternative measurement could be the anticipated probability distribution of a volatile outcome e.g., inflation visualized like the well-known "fan chart" (Bank of England, 2015).

In Stage 3, around these within-sample forecasts from various methods, we calculate Theil′s U values and forecast intervals to compare and contrast their accuracy, precision, and certainty. We then select the better-performing models (in terms of both how expenditures are measured and the projection models) based on such comparisons.

In Stage 4 we apply the selected models to the full set of data, to re-estimate parameters, and based on these estimates, conduct out-of-sample forecasts for 2015–2025. If such forecasts were logarithmized, then we transform them back to their original scale using and comparing two alternative ways of exponentiation[10]. One way is to directly exponentiate the logged estimated mean of expenditure y for the unlogged mean of forecast e^y; and the logged estimated standard error of the forecast σ for the unlogged standard error of the forecast $\sqrt{e^{\sigma^2}}$. The alternative way is to exponentiate both the logged estimated mean and the logged variance for the unlogged mean of forecast $e^{y+(\sigma^2)/2}$, and for the unlogged standard error of forecast $\sqrt{e^{2y+\sigma^2}(e^{\sigma^2}-1)}$ (Lin, 2012).

To further reduce the, randomness of individual forecasts obtained from each preferred model, in Stage 5, we combine the out-of-sample forecasts across models. The literature shows that either simple arithmetic average or complicated combination (different weights) of forecasts considerably increase accuracy and precision (Clemen, 1989; Elliott and Timmermann, 2008), and reduce uncertainty (Makridakis and Winkler, 1983; Makridakis *et al.*, 1998), compared to one individual forecast. The most commonly used combination is to take the average.

[10]To get a sense that how scaling down (i.e., logged) of the data (i.e., expenditures) can distort the forecast of the scaled-up data (i.e., exponentiate the logged expenditures).

5 Data

All health expenditure and macroeconomic statistics are from the OECD System of Health Accounts (SHA) database. The data are collected annually for 34 member countries. Although the length of the period varies slightly among countries, the 1960-2008 time series is much longer than any other international databases available. Therefore, we apply SHA to both time series and panel data models. For the completeness of available data for most countries, we choose types of health expenditures mainly classified by financing agents, and focus specifically on THE and PHE (classified by financing sources). These two types of expenditures are of considerable policy interest (OECD, 2011). We include countries and years that together can provide the largest number of observations and create strongly balanced panels. Specifically, we choose 20 countries from 1972 to 2008 (T=37) for THE and thus have 740 observations. The sample includes 18 countries over 1980-2008 (T=29) to provide 522 observations for PHE (Table 2). Three alternative datasets are also identified: World Health Organization (WHO) National Health Accounts (NHA) from 1995 to 2009; World Bank (WB) World Development Index (WDI) data for 1970–2009; International Monetary Fund (IMF) Government Finance Statistics (GFS) incompletely covering 2002–2008. These series, unfortunately, are too short to provide stable parameter estimates by time series approaches, we only use them to assess the quality of data from OECD-SHA. All estimation was performed using STATA 12.0.

6 Results

6.1 Comparison of Within-Sample Forecast Accuracy and Precision

In the interest of international comparison, $ refers to the International Dollar, which is the current national currency unit adjusted by PPP. We choose countries (N) and years (T) that together provide the largest number of observations and strongly balanced panels. So the N and T differ for the panels of THE and PHE, and the expenditure amounts cannot be compared between the two panels. This is because OECD countries differ largely in the total population. Table 2 reports the descriptive statistics of variables of 20 countries (N=20) during 1972–2008 (T=37) for forecasting THE. The mean of THE is $12,430 million in total and $1,210 per capita. The mean of GDP is $66,043 million in total and $15,975 per capita. The mean of the total population is 10,392 in thousands. The female and the elderly respectively account for about 50.87% and 13.22% of the population on average for these OECD countries.

Table 2 also reports the descriptive statistics of variables of 18 countries (N=18) during 1980–2008 (T=29) for forecasting PHE. The mean of PHE is $12,539 million in total and $1,077 per capita. The mean of GGR is $133,119 million in total, and $9,119 per capita. The mean of the total population is 11,610 in thousands. The female and the elderly respectively account for about 50.80% and 13.01% of the population on average for these OECD countries.

Tables 3 and 4 provide the indicators of forecast performance of different measurements of Canadian THE and PHE as well as the performance of statistical smoothing models versus econometric time series models. We only conduct comparisons of expenditure measurements using Canadian series data since each country has one time series, and here we are interested in how (time) series models perform differently

rather than how different countries' data perform differently. The fact that all Theil′s U values for statistical and time series data models are smaller than 1 indicates that these models are all better predictors than the naïve model. We report the results of four comparisons: (a) total vs. per capita measures, (b) logged vs. unlogged scales, (c) alternative currency units of expenditures, and (d) one statistical smoothing and three time series models.

First, for all time series models except VAR in only one case, total measures overall produce smaller Theil′s U values than per capita measures. For example, in the last column of Table 3, the Theil′s U value 0.0043 for totals in the first row of the first panel is slightly smaller than 0.0064 for per capita in the first row of the third panel, conditional on logged scale and international dollars produced by the VAR model.

Second, the "logged" scale provides smaller Theil′s U values than the "unlogged" scale. For instance, comparing the first rows of the first and third panels with those of the second and fourth panels, the Theil′s U values 0.0043 and 0.0064 for logged totals and per capita respectively are smaller than 0.1502 and 0.1311 for the unlogged counterparts, in international dollars produced by the VAR model. Therefore, we prefer "logged total" for subsequent out-of-sample forecasts.

Third, other things equal, the choice among three currency units differs across models. International dollars at current PPP usually outperforms in the VAR model. For example, in the first row of Table 3, comparing the last column with the ninth and eleventh columns, the U value 0.0043 produced by VAR is smaller than U values 0.0072 and 0.0204 respectively by ARIMA-dynamic and double exponential (DE) models for logged totals. Whereas NCU adjusted by GDP deator in the middle often produces smaller U values in other time series models than VAR (e.g., 0.0069 provided by VAR is larger than 0.0009 and 0.0015 respectively by ARIMA-static and dynamic models for logged totals). Both NCU adjusted by PPP and by GDP deflator are better than NCU in nominal terms. We choose international dollars for out-of-sample forecast

throughout this paper to facilitate international comparison using OECD data and because projections of inflation provided by financial forecasters for the out-of-sample forecast may later introduce extra error to forecasts using GDP-deflator.

Fourth, comparing three groups of performance indicators, Table 3 shows that, as expected, both U and the relative measures — MAPE and RMSPE — are less sensitive to the scale of PHE series than the absolute measures. Although the numerical values of certain indicators change substantially across models, the order of model performance based on each of these indicators, conditional on data scale and currency unit, is stable. That is, ARIMA-static (AS) $<$[11] VAR $<$ ARIMA-dynamic (AD) $<$ double exponential smoothing (DE) for logged scale, both totals and per capita, in IntD. Table 4 shows similar results for THE except that VAR seems on par with double exponential smoothing. That is, ARIMA-static $<$ DE$<$ VAR $<$ ARIMA-dynamic for logged totals; and ARIMA-static $<$ VAR $<$ DE $<$ ARIMA-dynamic for logged per capita.

Tables 5 and 6 report results for THE and PHE of three further comparisons: (a) logged total vs. logged per capita, (b) logged vs. exponentiated in two ways, and (c) time series vs. panel data models using international dollars[12]. The first panel shows that the Theil′s U values for all models are smaller than 1, which suggests all are better than the naïve model. Comparing the first and third rows of each panel reporting Theil′s U, MAPE, and RMSPE values, we find that logged totals again have smaller values than logged per capita. The static panel data model with FE estimator (SP-FE) seems especially sensitive to the two alternative ways of exponentiation. For example, comparing the first rows of the fifth and eighth numerical columns in Table 5, the U value of the static panel-FE model for THE increases from 0.0051 to 0.0442

[11] $<$ represents less than in terms of the numerical value of U-statistic produced by models.

[12] Because soon we will provide and rank out-of-sample forecasts of OECD countries, the within-sample comparisons of measures and models chosen based on Canadian series should use the currency unit that allows international comparison.

under the strong assumption that the standard error of forecast equals zero, but explodes to 0.9336 when this strict assumption is relaxed. Because 0.9336 approaches 1, the static panel-FE model becomes the worst among projection models and not much better than the naïve model. In comparison, smoothing and time series models in the first through fourth numerical columns overall outperform panel data models for both logged and exponentiated expenditures.

Finally, comparing the first through seventh numerical columns in Table 5, for THE exponentiated without the assumption of zero estimated error[13], the order of performance based on Theil′s U follows as ARIMA-static < DE < VAR < Dynamic panel-xtabond[14] estimator = Dynamic panel-xtdpd[15] estimator < ARIMA-dynamic < Static panel-FE estimator. Table 6 shows that PHE has similar results. However, the double-exponential smoothing model performs worse, which may suggest its sensitivity to the length of initialization set as the first two-thirds of past data[16]. What's more, although the three estimators for the dynamic panel model all passed specification tests, only the Arellano and Bover/Blundell and Bond estimators outperform time series models.

6.2 Combined Out-of-Sample Point Forecasts, and Forecast Intervals

We report country-specific expenditures (Table 7) and ranks (Table 8) of THE and PHE in international dollars per capita to facilitate cross-country comparison. Table 7 reports both the point forecast and the upper and lower bounds[17] of the forecast

[13]To get a sense that how scaling down (i.e., logged) of the data (i.e., expenditures) can distort the forecast of the scaled-up data (i.e., exponentiate the logged expenditures).
[14]xtabond represents Arellano and Bond estimator.
[15]xtdpd refers an alternative estimator to Arellano and Bond, as well as Arellano and Bover/Blundell and Bond system estimators.
[16]This is the tradition of the forecasting literature.
[17]Defined in the confidence interval introduced in section 3.

for each country in 2015, 2020, and 2025. To facilitate comparison, Table 8 presents the results ranked by point forecast expenditures.

To compare the point forecasts of most interest, Table 8 shows that in 2015 the US is forecast to spend the largest amount of THE per capita — up to $10,413 — about 5–6 times greater than the point forecast for Turkey. Following the US, European countries such as Norway (NOR), the Netherlands (NLD), and Germany (DEU) consistently rank in the top 5–10 throughout the forecast horizon, with a level of $5,000–9,000. Among them, the Netherlands climbs much faster than others, replaces the US at $15,816 in 2020, and stays on top at $28,488 through to 2025. Others, such as Iceland (ISL), conversely, are forecast to decrease THE down to the bottom of the rank in the medium run. English-speaking countries except the US consistently locate in the middle. Specifically, Canada is forecast to increase THE to $6,101 in 2015, and up to $8,347 and $11,444 in 2020 and 2025 respectively, whereas Australia spends slightly higher (at $4,884) than GBR (at $4,857) in 2015, but lower (at $8,148) than GBR (at $9,036) in 2025.

PHE per capita forecast differs from THE. Norway, followed by the US, is forecast to become the largest public financier of health care at $6,662 and $12,835 in 2015 and 2025. Whereas the US' PHE is forecast to reach at $11,046 in 2025. The Netherlands' PHE, similar to its THE, rises much faster again than that of other countries up to $10,131 in 2025 and ranks third highest in absolute amount. Other English-speaking countries again remain in the middle. Canada's and Britain's PHE both rise slightly toward $6,000 and $8,500 in 2020 and 2025, higher than Australia's PHE of about $4,600 and $6,200 for the same years.

The ratio of PHE to THE, however, shows a different picture from the absolute amounts noted above. In the rank of per capita PHE as the proportion of per capita THE, GBR is forecast to have the persistently largest public share (i.e., near 90%) of THE until 2020. European countries consistently rank in the top 5-10, while the

Netherlands is forecast to have the second lowest public share[18]. Meanwhile, the US' PHE accounts for half of its THE and incrementally grows to 53%, 59%, and 65% of THE in 2015, 2020, and 2025. Canada's PHE accounts for a large share (i.e., 80%)[19] and remains slightly lower in the future around 74%. Japan's PHE share is expected to climb from 84% in 2015 up to 92% in 2025. Whereas Portugal's drops from 53% in 2015 to 34% in 2025 toward the bottom of the rank.

Figures 1–4 show that forecasts in total differ slightly from those in per capita above. It is the two large economies of Japan and Germany, rather than the Netherlands, who are expected to follow the US in spending the most on health care. Such patterns are also observed in PHE forecasts in absolute total terms. The US, Canada, and European countries such as Iceland show apparent reductions in THE (Figures 1 and 2) and PHE (Figures 3 and 4) in 2009 possibly because their 2008 data already reflected an economic slow-down. Whereas Australia's THE appears relatively smooth and shows a slightly increasing curve whose positive slope is smaller than that of Japan. South Korea, as one of the emerging Asian economies, possesses the steepest upward curve. Iceland's PHE, in particular, is flatter than its THE curve, whereas the Britain's and the US' PHEs show steeper downward curves in 2009 than their THEs. These observations coincide with the stringent budget constraints of these countries since the middle of 2008.

[18] About 51% of its THE in 2015, and quickly shrinks public support down to 43% in 2020 and 36%, lower than the US', in 2025.

[19] This is higher than Canadian historical records at an average 70% during 2000–2012. Recall this is the forecast combined from four smoothing and time series models. The individual forecast of PHE/THE provided by individual model DE, HW, AD, and VAR respectively is 74%, 97%, 71%, and 80%. Obviously, the final combined 80% comes from the VAR and especially the HW model, which corresponds to HW's high sensitivity to the end points of data used to forecast.

6.3 Data Discrepancy

Working with these international databases has identified a number of potential quality issues. Below we illustrate some particularly large[20] and systematic discrepancies in comparison of the same data across different databases. Table 9 shows that PHE provided by the OECD and WHO databases for Switzerland are both 24 billion NCU higher than that provided by the IMF database in 2008, by a factor of 3. Similarly, OECD and WHO database derived PHEs for Denmark are respectively 16 and 7 NCU billion higher, on average, over 2002–2009. Out-of-pocket payment (OOP) provided by the OECD database for Japan, compared to that provided by the WHO database, is on average 73 billion NCU lower, ranging from 671 billion lower in 1998 to 852 billion higher in 2006. GDP series for South Korea provided by the OECD and UN databases are consistent, but both are 14,956 billion NCU higher than that provided by the WB database. WHO data for this series is even higher (i.e., 19,976 billion) on average over 1995–1999, but all four databases for this series suddenly harmonize beginning in 2000. The PHE as a percentage of GGE for Switzerland offered by the OECD or WB databases is about 14 percentage points higher than that provided by the IMF database for 2007. Such a percentage offered by the OECD or WB databases, compared to that for the WHO database, is, on average, 10 percentage points higher for Chile and lower for Finland. On the other hand, OECD, WHO, WB data only harmonize for South Korea after the year 2000, for Chile after 2003, and for Denmark before 2003. This suggests possible sources of data discrepancy including data management changes and exogenous events occurring in these countries during these years.

[20]Large discrepancy in international databases means the correlation coefficient between two amounts of the same variable in two databases is below 0.5.

7 Discussion and Conclusion

Our results indicate that, contrary to Getzen and Poullier (1992), complicated econometric (esp., static) panel data models perform worse than simpler statistical (e.g., smoothing) and econometric time series models for forecasting. This finding is inconsistent with the early findings by Getzen and Poullier (1992) and responds to their call for forecasting based on more years (nearly 20, in our case) of data. This finding is, however, consistent with empirical findings from the broader literature on economic forecasting (Armstrong, 1978; McNees, 1986). Therefore, to forecast, more complicated econometric models — except VAR — generally do not do better than simpler ones such as exponential smoothing. This implies that a (usually complex) model that better fits historical data does not guarantee accurate and precise post-sample forecasts. This paper also finds that some methods perform more accurately for short horizons (e.g., exponential smoothing) while others are more appropriate for medium horizons (e.g., ARIMA, VAR). This is also consistent with studies from the broader literature on forecasts (Fildes and Makridakis, 1995; Fildes *et al.*, 1998).

The recent literature review (Astolfi *et al.*, 2012) on health expenditure forecasting suggests that with better computing power and more refined data, the future of forecasting is complicated micro models. But micro panel data models can forecast worse than time series and smoothing models. Our paper confirms this. Astolfi *et al.* (2012) also suggest structural models (e.g., CGE) for forecasting that specify explicit causal-effect mechanisms among variables. In comparison, the models used in this paper are reduced-form and hence "atheoretical", which is fine since our objective is to obtain accurate and precise forecasts of future spending rather than to understand past determinants. It is important to carefully distinguish the objective of simply getting accurate forecasts of future spending from modeling from that of understanding

determinants of expenditure growth. Arriving at such an understanding of the determinants of expenditure growth requires considerably more data and effort, and may still do worse with pure forecasting. Our paper confirms this. Thus, different kinds of projection models may work better in different situations and respond to different demands from policy makers. For instance, theoretical models are more suitable for "if ... then" policy questions. For health policy makers in the Ministry of Health, forecasts based on a health-sector-only structural model would be enough. Whereas for those who are responsible for more general issues (e.g., analysts in Ministry of Finance) across all major sectors (not only health but also education, manufacturing, etc.), a CGE model is demanded. Unfortunately, a well-agreed theoretical model for health expenditure at the macro level has not yet been well established, though there are prototypes of CGE models (Astolfi *et al.*, 2012). This might be why articles on health expenditure forecasts are few, although cost estimates are many. This inquiry urges, as Gerdtham and Jonsson (2000) called for more than a decade ago, both macroeconomic theories and macro structural models for health expenditure (Heckman, 2000; Carnot *et al.*, 2011).

Our paper also contributes to the call made by Astolfi *et al.* (2012) for more rigorous methods, for more transparency, and for better assessment of performance. Policy makers could choose time series models that treat the system as a "black box" and do not attempt to discover the factors affecting its behavior. Univariate time series methods also suit when the main concern is to forecast what will happen to expenditures, not why or how it happens. If the latter is important, explanatory forecasts based on multivariate models from time series data such as VAR or panel data will be preferred. Elliott and Timmermann (2008), however, emphasize that time series models can be unstable so that one cannot rely on the same dominant model in different historical samples. Hendry and Hubrich (2012) similarly conclude that after introducing the uncertainty of a variable such as the Consumer Price Index

(CPI), aggregate forecast using aggregate data performs less accurately than that using summarized disaggregate information. However, these are not concerns in terms of our objective — to forecast health expenditures, which are less volatile and hence less uncertain than variables like stock price and CPI. Nonetheless, we recommend that health expenditure forecasters who have chosen one kind of time series model continually compare its performance with that of other kinds of time series models. Although we argue Theil′s U statistic is the superior measure of the performance of projection models, we are aware that judgment of models might differ when using alternative indicators (Makridakis *et al.*, 1993). Characteristics such as randomness and frequency of data matter for forecasting as well. For annual HE data like ours in which the trend dominates cyclical fluctuations, and little randomness is present, we recommend Holt-Winters non-seasonal and its alternative — double exponential smoothing, and time series models. Finally, we are aware that the international databases have been updated and regularly revised so that particular patterns of the data might be susceptible to such revisions and lead to models' misspecification. We recommend using the most recently updated data.

Budgeting and strategic planning for the short and medium run, and fiscal sustainability in the long run can only be achieved with sound forecast modeling, monitoring, and modifying (Makridakis *et al.*, 1998). Forecasts provided by either the academia or the policy arena need to be updated for periodic changes and tracked on records of performance for the purposes of modifying and improving methods in both theory and practice.[21] Health ministers usually favor higher growths of the public health budget, but finance ministers prefer lower. No matter how objectively health expenditure forecasting is improved, decisions concerning allocation of monetary resources

[21]Recall that it took decades for Samuelson to withdraw his previous prediction from his textbook that Soviet GNP would exceed that of the United States by as early as 1984 or perhaps by as late as 1997 (Samuelson, 1980). "The future will prove that anything could be wrong" reflected by Thomas Getzen.

to the health sector inevitably involve the human judgment that may be inuenced by bargaining power and political considerations hence have little to do with objectivity.

Table 2: Variable statistical summary of panel data for OECD countries

Total health expenditure (THE), 20 countries, 1972-2008				Public health expenditure (PHE), 18 countries, 1980-2008			
Variable	Observations	Mean	Std. Dev.	Variable	Observations	Mean	Std. Dev.
THE_{MIntD}	739	12430.40	6.68	PHE_{MIntD}	521	12539.01	6.13
THE_{PCIntD}	716	1210.03	2.30	PHE_{PCIntD}	506	1076.75	2.23
GDP_{MIntD}	740	66042.66	5.85	GGR_{MIntD}	392	133119.17	4.89
GDP_{PCIntD}	740	15975.32	1.94	GGR_{PCIntD}	392	9118.86	1.88
POPTOT	740	10392.09	5.25	POPTOT	522	11609.74	5.88
POPFE	740	50.87	1.01	POPFE	522	50.80	1.01
POPELDER	740	13.22	1.21	POPELDER	522	13.01	1.30

Notes: According to definitions from OECD et al. (2011), THE refers to total spending executed on health goods and services whose primary purposes include maintenance, restoration or enhancement of health (PHE). GGR is general revenues of all levels of government. POPTOT refers to total population size, in thousands. POPFE is the proportion of the female population. POPELDER is the proportion of the elderly aged at 65 and above.
Monetary variables are measured in millions of International Dollars, IntD, in total and per capita respectively. MIntD refers to million international dollars (IntD). PCIntD means per capita IntD.
The observations of THE and PHE differ, because we include countries and years that together can provide the largest number of observations and create strongly balanced panels.

Table 3: Comparison of forecast performance of measurements of health expenditures using alternative indicators, Public health expenditure (PHE), Canada, within test set 1997–2008

		NCU, Current Price				NCU, GDP deflator				IntD, Current PPP			
Measure of series	Performance indicator	DE	AS	AD	VAR	DE	AS	AD	VAR	DE	AS	AD	VAR
Total, logged	Theil′s U	0.0212	0.0019	0.0108	0.0192	0.0176	0.0009	0.0015	0.0069	0.0204	0.0012	0.0072	0.0043
	MAE	0.4083	0.0369	0.2322	0.3423	0.3453	0.0132	0.0306	0.1255	0.3873	0.0226	0.1507	0.0739
	MSE	0.2209	0.0019	0.0655	0.1807	0.1514	0.0004	0.0012	0.0235	0.1971	0.0007	0.0275	0.0090
	MAPE	3.5765	0.3289	2.0384	2.9930	3.0543	0.1185	0.2717	1.1089	3.4534	0.2055	1.3464	0.6588
	RMSPE	4.0928	0.3857	2.2330	3.6950	3.4298	0.1718	0.3019	1.3505	3.9342	0.2339	1.4740	0.8386
Total, unlogged	Theil′s U	0.2528	0.0094	0.2645	0.1322	0.2295	0.0098	0.2783	0.1306	0.2403	0.0140	0.2481	0.1502
	MAE	28925	1377.3	29955	161501	25562	1128.5	29648	15969	22816	1871.3	23392	15081
	MSE	1.2e+09	24909851	3e+09	3.9e+08	8.5e+08	22727821	2.2e+09	3.2e+08	7.4e+08	37342407	7.8e+08	3.3e+08
	MAPE	31.698	1.9085	32.787	17.500	31.485	1.7140	36.335	19.821	30.542	2.9058	31.285	20.080
	RMSPE	35.449	2.3236	36.742	19.935	34.656	2.3974	40.519	21.423	34.028	3.0232	34.910	22.567
Per capita, logged	Theil′s U	0.0313	0.0023	0.0138	0.0070	0.0255	0.0012	0.0031	0.0082	0.0301	0.0017	0.0094	0.0064
	MAE	0.4145	0.0294	0.2090	0.0799	0.3445	0.0124	0.0453	0.0997	0.3917	0.0221	0.1377	0.0747
	MSE	0.2272	0.0013	0.0529	0.0117	0.1505	0.0004	0.0024	0.0160	0.2013	0.0006	0.0230	0.0095
	MAPE	5.2156	0.3797	2.6376	1.0019	4.3949	0.1609	0.5799	1.2689	5.0571	0.2923	1.7829	0.9620
	RMSPE	5.9548	0.4622	2.8804	1.3449	4.9320	0.2481	0.6264	1.6054	5.7479	0.3371	1.9475	1.2419
Per capita, unlogged	Theil′s U	0.2516	0.0093	0.2734	0.1041	0.2338	0.0081	0.2768	0.1199	0.2382	0.0144	0.2591	0.1311
	MAE	912.82	42.177	973.78	413.52	824.82	28.891	939.30	473.10	718.11	59.427	766.76	428.84
	MSE	1167165	2382.6	1341489	250206	876669	1536.7	1167010	271973	718434	3853.2	827665	255422
	MAPE	32.215	1.8138	34.291	14.493	32.576	1.2822	36.969	18.852	30.955	2.9277	32.976	18.513
	RMSPE	35.991	2.2243	38.474	16.390	35.917	1.2822	41.246	20.242	34.447	3.0860	36.866	20.514

Notes: NCU refers to national currency unit. DE refers to Double-Exponential smoothing model; AS refers to ARIMA model with static forecast; AD refers to ARIMA model with dynamic forecast. VAR refers to VAR model. RMSPE means the square root of MSPE.

Table 4: Comparison of forecast performance of measurements of health expenditures using alternative indicators, Total health expenditure (THE), Canada, within test set 1997–2008

		NCU, Current Price				NCU, GDP deflator				IntD, Current PPP			
Measure of series	Performance indicator	DE	AS	AD	VAR	DE	AS	AD	VAR	DE	AS	AD	VAR
Total, logged	Theil′s U	0.0108	0.0013	0.0209	0.0128	0.0167	0.0006	0.0024	0.0106	0.0050	0.0014	0.0180	0.0131
	MAE	0.1905	0.0240	0.4577	0.2658	0.3400	0.0105	0.0496	0.2246	0.0557	0.0266	0.3856	0.2710
	MSE	0.0618	0.0009	0.2430	0.0901	0.1446	0.0002	0.0030	0.0592	0.0101	0.0010	0.1720	0.0910
	MAPE	1.6175	0.2092	3.9280	2.2742	2.9164	0.0911	0.4268	1.9293	0.5131	0.2353	3.3665	2.3606
	RMSPE	2.0981	0.2610	4.2051	2.5489	3.2519	0.1234	0.4695	2.0843	0.8750	0.2839	3.5995	2.6096
Total, unlogged	Theil′s U	0.1644	0.0075	0.1387	0.1240	0.2067	0.0129	0.0963	0.1578	0.1406	0.0119	0.1268	0.1085
	MAE	23595	1201.1	20290	18072	33491	2528.4	16911	26986	16773	1800.0	15264	12952
	MSE	1.1e+09	2819968	7.8e+08	6.4e+08	1.5e+09	7789656	3.8e+08	9.1e+08	5.4e+08	4803428	4.5e+08	3.4e+08
	MAPE	17.903	1.1535	15.417	13.687	28.971	2.3897	14.586	23.514	15.516	2.0579	14.132	11.950
	RMSPE	22.944	1.7091	19.701	17.665	31.773	2.6855	16.147	25.361	19.885	2.4718	18.089	15.580
Per capita, logged	Theil′s U	0.0143	0.0016	0.0261	0.0035	0.0240	0.0008	0.0043	0.0103	0.0100	0.0018	0.0226	0.0024
	MAE	0.1761	0.0188	0.4062	0.0459	0.3411	0.0099	0.0635	0.1424	0.1204	0.0246	0.3430	0.0318
	MSE	0.0525	0.0007	0.1889	0.0032	0.1454	0.0002	0.0048	0.0274	0.0246	0.0008	0.1343	0.0014
	MAPE	2.1205	0.2355	4.9572	0.5635	4.1646	0.1219	0.7767	1.7387	1.4870	0.3127	4.2899	0.4022
	RMSPE	2.7381	0.3244	5.2680	0.6858	4.6395	0.1698	0.8475	2.0124	1.9166	0.3699	4.5529	0.4820
Per capita, unlogged	Theil′s U	0.1539	0.0078	0.1350	0.0998	0.2097	0.0073	0.0765	0.1369	0.1277	0.0110	0.0931	0.0894
	MAE	713.09	37.802	635.27	476.49	1077.0	40.151	436.87	759.895	493.17	50.430	369.568	348.04
	MSE	934016	2990.9	738657	428893	1480130	2520.1	243044	700197	451902	4072.3	252996	236671
	MAPE	17.472	1.1164	15.590	11.695	29.870	1.2281	12.115	21.241	14.743	1.8107	11.096	10.404
	RMSPE	22.309	1.6775	19.836	14.947	32.790	1.5877	13.338	22.793	18.802	2.2476	14.053	13.468

Notes: NCU refers to national currency unit. DE refers to Double-Exponential smoothing model; AS refers to ARIMA model with static forecast; AD refers to ARIMA model with dynamic forecast. VAR refers to VAR model.

Table 5: Comparison of forecast performance of time series models versus panel data models using alternative indicators, THE, logged total and per capita expenditures adjusted by PPP, Canada, within test set 1997–2008

	Series Measure	DE	AS	AD	VAR	SP-FE-exp1	DP1-exp1	DP3-exp1	SP-FE-exp2	DP1-exp2	DP3-exp2
Theil′s U	total logged	0.0050	0.0014	0.0180	0.0131	0.0051	0.0149	0.0149	0.0051	0.0149	0.0149
	total exponentiated	0.1071	0.0142	0.2423	0.1831	0.0442	0.2055	0.2055	0.9336	0.2333	0.2333
	per cap logged	0.0100	0.0018	0.0226	0.0024	0.0072	0.0174	0.0174	0.0072	0.0174	0.0174
	per cap exponentiated	0.0925	0.0129	0.2092	0.0190	0.0632	0.1425	0.1425	0.9470	0.1583	0.1583
MAE	total logged	0.0557	0.02657	0.3856	0.2710	0.0839	0.2590	0.2590	0.0839	0.2590	0.2590
	total exponentiated	13086	2288.1	48404	32724	11860	60483	60483	3847245	60545	60545
	per cap logged	0.1204	0.0246	0.3430	0.0318	0.0872	0.2323	0.2323	0.0872	0.2323	0.2323
	per cap exponentiated	364.57	67.186	1301.1	95.277	272.10	671.29	671.29	101480	671.30	671.30
MSE	total logged	0.0101	0.0010	0.1720	0.0910	0.0114	0.0969	0.0969	0.0114	0.0969	0.0969
	total exponentiated	3.3e+08	6953178	3.4e+09	1.7e+09	1.4e+09	4.3e+10	4.3e+10	1.3e+14	4.3e+10	4.3e+10
	per cap logged	0.0246	0.0008	0.1343	0.0014	0.0129	0.0751	0.0751	0.0129	0.0751	0.0751
	per cap exponentiated	251377	5660.5	2316260	12114	161626	805622	805622	1.2e+10	806296	806296
MAPE	total logged	0.5131	0.2353	3.3665	2.3606	0.8437	2.7340	2.7340	0.8437	2.7340	2.734
	total exponentiated	12.144	2.7184	48.767	32.268	8.5896	24.482	24.482	2668.1	24.481	24.481
	per cap logged	1.4870	0.3127	4.2899	0.4022	1.0991	2.9355	2.9355	1.0991	2.9355	2.9355
	per cap exponentiated	10.937	2.5028	42.107	3.4360	9.0135	22.228	22.228	3474.8	22.226	22.226
RMSPE	total logged	0.8750	0.2839	3.5995	2.6096	1.1251	3.6799	3.6799	1.1251	3.6799	3.6799
	total exponentiated	15.502	3.2731	53.495	36.604	11.172	28.663	28.663	2684.3	28.668	28.668
	per cap logged	1.9166	0.3699	4.5529	0.4820	1.4253	3.4520	3.4520	1.4253	3.4520	3.4520
	per cap exponentiated	13.951	2.9514	45.692	3.9776	12.073	25.690	25.690	3498.0	25.694	25.694

Notes: SP-FE denotes static panel data model using fixed-effect estimator. DP1–3 denote -xtabond and -xtdpd respectively, which represent dynamic panel data model using Arellano-Bond and Arellano-Bover/Blundell-Bond estimators.

exp1 refers to one way to exponentiate logged expenditures using eyf=e^{yf} for the mean of forecast, and eyfstdf=$\sqrt{e^{yfstdf^2}}$ for the standard error of forecast. exp2 refers to alternative and preferable way to exponentiate logged expenditures using eyf=$e^{yf+(yfstdf^2)/2}$ for the mean of forecast, and eyfstdf=$\sqrt{e^{2yf+yfstdf^2}(e^{yfstdf^2}-1)}$ for the standard error of forecast (Lin, 2012).

Table 6: Comparison of forecast performance of time series models versus panel data models using alternative indicators, PHE, logged total and per capita expenditures adjusted by PPP, Canada, within test set 1997–2008

	Series Measure	DE	AS	AD	VAR	SP-FE-exp1	DP1-exp1	DP2-exp1	DP3-exp1	SP-FE-exp2	DP1-exp2	DP2-exp2	DP3-exp2
Theil′s U	total logged	0.0204	0.0012	0.0072	0.0043	0.0065	0.0089	0.0043	0.0143	0.0065	0.0089	0.0043	0.0143
	total exponentiated	0.2397	0.0109	0.0465	0.0577	0.0605	0.0833	0.0406	0.1327	0.4783	0.0915	0.0437	0.1460
	per cap logged	0.0301	0.0017	0.0094	0.0064	0.0087	0.0132	0.0051	0.0191	0.0087	0.0132	0.0051	0.0191
	per cap exponentiated	0.2364	0.0111	0.0878	0.0328	0.0653	0.0958	0.0369	0.1319	0.5967	0.1041	0.0401	0.1432
MAE	total logged	0.3873	0.0226	0.1507	0.0739	0.1081	0.1411	0.0687	0.2317	0.1081	0.1411	0.0687	0.2317
	total exponentiated	22766	1427.5	12397	5833.8	9035.9	13508	5515.5	22306	136008	13563	5511.6	22369
	per cap logged	0.3917	0.0221	0.1377	0.0747	0.1071	0.1597	0.0584	0.2313	0.1071	0.1597	0.0584	0.2313
	per cap exponentiated	713.42	43.931	348.86	168.91	224.78	305.76	112.72	427.32	4427.4	305.70	112.63	427.25
MSE	total logged	0.1971	0.0007	0.0275	0.0090	0.0176	0.0326	0.0077	0.0849	0.0176	0.0326	0.0077	0.0849
	total exponentiated	7.4e+08	2375067	2.2e+08	59249140	6.2e+08	1.3e+09	2.5e+08	3.7e+09	8.8e+10	1.3e+09	2.5e+08	3.7e+09
	per cap logged	0.2013	0.0006	0.0230	0.0095	0.0177	0.0390	0.0057	0.0814	0.0177	0.0390	0.0057	0.0814
	per cap exponentiated	709565	2311.7	170903	53974	93128	156303	23505	293149	26444234	156227	23414	293096
MAPE	total logged	3.4534	0.2055	1.3464	0.6588	1.1292	1.5800	0.7474	2.6033	1.1292	1.5800	0.7474	2.6033
	total exponentiated	30.473	2.2880	16.606	7.7000	10.583	13.373	6.7411	21.228	201.06	13.373	6.7444	21.231
	per cap logged	5.0571	0.2923	1.7829	0.9620	1.4271	2.1299	0.7820	3.0847	1.4271	2.1299	0.7820	3.0847
	per cap exponentiated	30.748	2.2332	15.056	7.1125	10.575	15.189	5.7323	21.339	287.71	15.189	5.7326	21.339
RMSPE	total logged	3.9342	0.2339	1.4740	0.8386	1.4633	2.2880	1.0361	3.7014	1.4633	2.2880	1.0361	3.7014
	total exponentiated	33.9577	2.6024	18.473	9.2605	13.128	16.369	8.4508	25.097	203.57	16.368	8.4488	25.101
	per cap logged	5.7479	0.3371	1.9475	1.2419	1.7965	2.6156	1.0134	3.7741	1.7965	2.6156	1.0134	3.7741
	per cap exponentiated	34.229	2.5665	16.713	9.0803	13.376	17.985	7.3081	24.781	290.73	17.986	7.3032	24.784

Notes: SP-FE denotes static panel data model using fixed-effect estimator. DP1,2,3 denote -xtabond, -xtdpdsys, and xtdpd respectively, which further represent dynamic panel data model using Arellano-Bond and Arellano-Bover/Blundell-Bond estimators.

exp1 refers to one way to exponentiate logged expenditures using eyf=e^{yf} for the mean of forecast, and eyfstdf=$\sqrt{e^{yfstdf^2}}$ for the standard error of forecast. exp2 refers to alternative and preferable way to exponentiate logged expenditures using eyf=$e^{yf+(yfstdf^2)/2}$ for the mean of forecast, and eyfstdf=$\sqrt{e^{2yf+yfstdf^2}(e^{yfstdf^2}-1)}$ for the standard error of forecast (Lin, 2012).

Table 7: Combined out-of-sample point forecasts and forecast intervals of per capita THE and PHE in International Dollars for OECD countries in 2015, 2020, and 2025

		THE			PHE				THE			PHE		
		2015	2020	2025	2015	2020	2025		2015	2020	2025	2015	2020	2025
	Upper	5380	6966	9090	3826	5031	6705		7751	10769	14770	4027	5452	7517
Australia	Point	4884	6292	8148	3427	4604	6231	Netherlands	9025	15816	28488	4596	6789	10131
	Lower	4596	6032	7968	2723	3611	4849		6102	8295	11307	3291	4509	6277
	Upper	7326	10099	13937	5548	7573	10404		4750	6110	7889	4228	5608	7864
Austria	Point	6076	8150	10969	4775	6453	8759	New Zealand	4290	5870	8069	3686	5381	7969
	Lower	5366	7432	10366	4169	5689	7841		3346	4344	5667	3057	4506	6318
	Upper	6278	8337	11064	5331	8745	14766		9780	14361	21017	7327	10446	15039
Canada	Point	6101	8347	11444	4911	6073	8426	Norway	7963	11120	15633	6662	9199	12835
	Lower	5155	6895	9279	3510	4383	5768		7390	10918	16336	6476	9136	13078
	Upper	7604	9281	12792	5824	7846	10534		5213	7750	11742	2773	3558	4369
Denmark	Point	6456	9015	12660	5376	7308	9910	Portugal	4224	6186	9168	2226	2669	3148
	Lower	5506	8339	11493	4622	5897	7451		3508	5265	8038	1836	1950	2017
	Upper	5001	6710	9034	3734	5194	7331		5339	8224	13027	4141	7361	11791
Finland	Point	4371	5757	7615	3239	4352	5904	Spain	4653	6855	10311	3495	5230	7861
	Lower	3851	5229	7162	2839	4056	5858		4322	6846	11155	2997	4408	7041
	Upper	6473	8666	11891	5033	7033	9988		5079	6341	8066	4368	5506	6838
Germany	Point	5878	7825	10548	4513	6031	8154	Sweden	4697	5848	7343	3889	4756	5767
	Lower	5387	7303	10134	4173	5610	7740		4210	5351	6900	3314	3766	4183
	Upper	5588	7404	9910	4452	5802	7762		2788	9704	4958	–	–	–
Iceland	Point	4458	5557	7025	3625	4442	5559	Turkey	1863	3149	5375	–	–	–
	Lower	4112	5598	7716	3238	4213	5641		1255	1462	3421	–	–	–
	Upper	5711	7213	8786	4918	6712	9370		5173	6781	9054	4468	6474	8706
Ireland	Point	5067	6776	9055	3820	4999	6668	UK	4857	6582	9036	4256	6072	8468
	Lower	2905	3524	4271	3228	4470	6393		3976	5215	7063	3534	4877	6536
	Upper	5295	7600	11147	4397	6747	10694		11566	15091	19770	5985	8609	12404
Japan	Point	4372	5984	8324	3683	5242	7678	US	10413	13287	17049	5536	7812	11046
	Lower	4127	6015	8943	3781	5830	9268		9579	12598	16773	5050	7330	10675
	Upper	3899	6208	9861	2270	3693	5710							
South Korea	Point	3276	5184	8206	2018	3429	5709							
	Lower	2730	4216	6492	1480	2348	3500							

Notes: Canada's PHE/THE point forecast in 2015 is 80%, which is higher than historical average as 70% during 2000–2012. Recall this is the forecast combined from four smoothing and time series models. The individual forecast of PHE/THE provided by individual model DE, HW, AD, and VAR respectively is 74%, 97%, 71%, and 80%. Obviously, the final combined 80% comes from the VAR and especially the HW model, which corresponds to HW's high sensitivity to the end points of data used to forecast.

Similarly, the Netherlands' forecasts are much larger than the US' forecasts, because that the individual forecasts from the statistical smoothing models especially HW model for the Netherlands are much larger than those for the US. Also, the Netherlands has a higher rate of growth in PHE because of their reforms in 2005–2006, which gets projected forward.

Table 8: Ranks of combined point forecasts from time series data models, per capita, International Dollars, OECD countries, 2015, 2020, and 2025

	THE						PHE						PHE/THE					
Rank	2015		2020		2025		2015		2020		2025		2015		2020		2025	
1	USA	10413	NLD	15816	NLD	28488	NOR	6662	NOR	9199	NOR	12835	GBR	0.88	GBR	0.92	NZL	0.99
2	NLD	9025	USA	13287	USA	17049	USA	5536	USA	7812	USA	11046	NZL	0.86	NZL	0.92	GBR	0.94
3	NOR	7963	NOR	11120	NOR	15633	DNK	5376	DNK	7308	NLD	10131	JPN	0.84	JPN	0.88	JPN	0.92
4	CHE	7079	CHE	9327	DNK	126604	CAN	4911	NLD	6789	DNK	9910	NOR	0.84	NOR	0.83	NOR	0.82
5	DNK	6456	DNK	9015	CHE	12281	AUT	4775	AUT	6453	AUT	8759	DNK	0.83	SWE	0.81	AUT	0.80
6	CAN	6101	CAN	8347	CAN	11444	NLD	4596	CAN	6073	GBR	8468	SWE	0.83	DNK	0.81	ISL	0.79
7	AUT	6076	AUT	8150	AUT	10969	DEU	4513	GBR	6072	CAN	8426	ISL	0.81	ISL	0.80	SWE	0.79
8	DEU	5878	DEU	7825	DEU	10548	GBR	4256	DEU	6031	DEU	8154	CAN	0.80	AUT	0.79	DNK	0.78
9	BEL	5490	BEL	7441	ESP	10311	SWE	3889	NZL	5381	NZL	7969	AUT	0.79	DEU	0.77	FIN	0.78
10	IRE	5067	ESP	6855	BEL	10186	IRE	3820	JPN	5242	ESP	7861	DEU	0.77	ESP	0.76	DEU	0.77
11	AUS	4884	IRE	6776	PRT	9168	NZL	3686	ESP	5230	JPN	7678	IRE	0.75	FIN	0.76	AUS	0.76
12	GBR	4857	GBR	6582	IRE	9055	JPN	3683	IRE	4999	IRE	6668	ESP	0.75	IRE	0.74	ESP	0.76
13	SWE	4697	AUS	6292	GBR	9036	ISL	3625	SWE	4756	AUS	6232	FIN	0.74	AUS	0.73	IRE	0.74
14	ESP	4653	PRT	6186	JPN	8324	ESP	3495	AUS	4604	FIN	5904	AUS	0.70	CAN	0.73	CAN	0.74
15	ISL	4458	JPN	5984	KOR	8206	AUS	3427	ISL	4442	SWE	5767	KOR	0.62	KOR	0.66	KOR	0.70
16	JPN	4372	NZL	5870	AUS	8148	FIN	3239	FIN	4352	KOR	5709	USA	0.53	USA	0.59	USA	0.65
17	FIN	4371	SWE	5848	NZL	8069	PRT	2226	KOR	3429	ISL	5559	PRT	0.53	PRT	0.43	NLD	0.36
18	NZL	4290	FIN	5757	FIN	7615	KOR	2018	PRT	2669	PRT	3148	NLD	0.51	NLD	0.43	PRT	0.34
19	PRT	4224	ISL	5557	SWE	7343	–	–	–	–	–	–	–	–	–	–	–	–
20	KOR	3276	KOR	5184	ISL	7025	–	–	–	–	–	–	–	–	–	–	–	–
21	TUR	1863	TUR	3149	TUR	5375	–	–	–	–	–	–	–	–	–	–	–	–

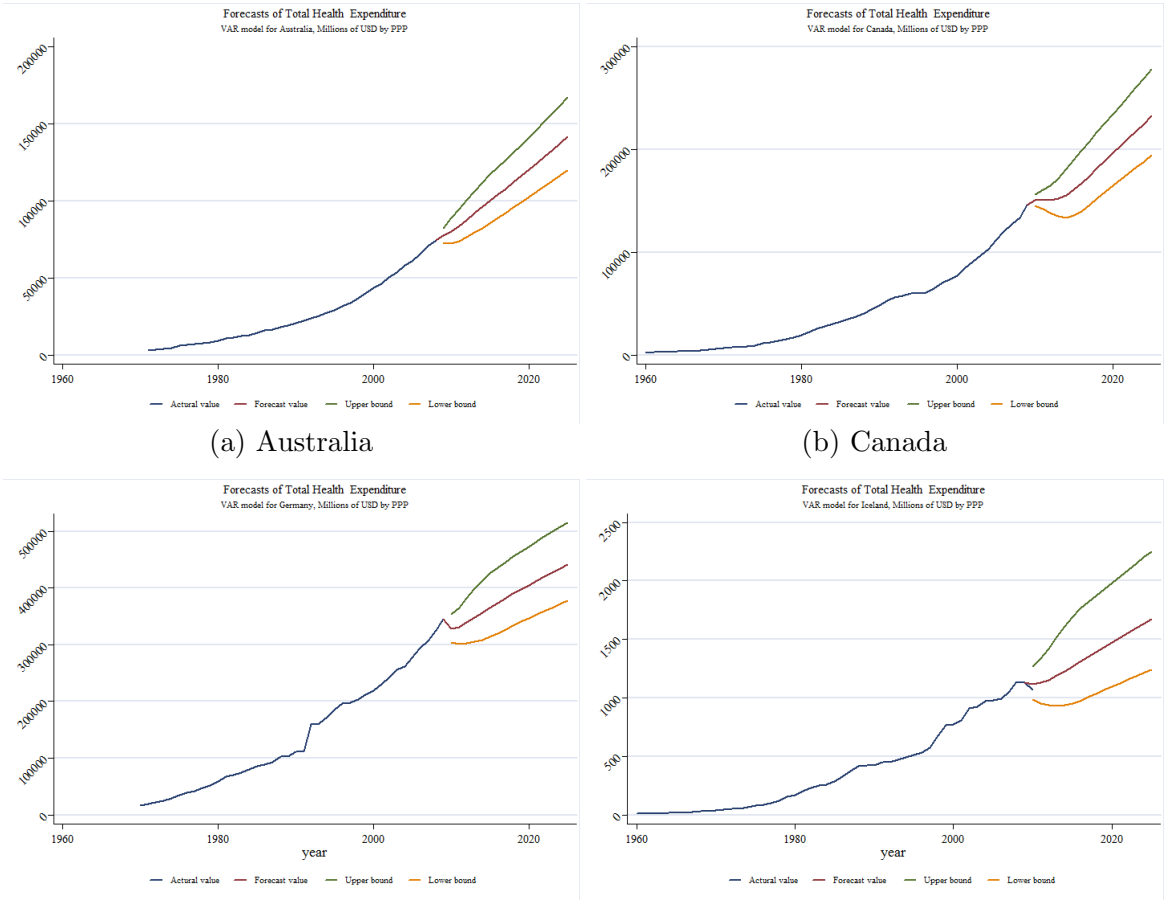
Notes: According to OECD country abbreviations, NLD—the Netherlands; NOR—Norway; CHE—Switzerland; DNK—Denmark; CAN—Canada; AUT—Austria; DEU—Germany; BEL—Belgium; IRE—Ireland; AUS—Australia; GBR—Great Britain; SWE—Sweden; ESP—Spain; ISL—Iceland; JPN—Japan; FIN—Finland; NZL—New Zealand; PRT—Portugal; KOR—South Korea; TUR—Turkey.

Table 9: Large and systematic discrepancies among international databases

PHE (Billions NCU)	Country	Switzerland	Denmark													
	Year	2008	2002	2003	2004	2005	2006	2007	2008	2009						
	OECD	35	108	113	120	128	137	143	151	162						
	WHO	34	100	104	110	116	125	132	138	149						
	IMF	11	96	99	105	111	119	127	135	–						
OOP (Billions NCU)		Japan														
	Year	1995	1996	1997	1998	1998	1999	2001	2002	2003	2004	2005	2006	2007	2008	
	OECD	4,775	5,020	5,468	5,775	5,774	5,959	5,977	6,099	6,540	6,543	6,350	7,059	6,807	6,774	
	WHO	5,198	5,544	6,059	6,446	6,432	6,524	6,517	6,763	6,079	6,112	5,843	6,207	6,115	6,098	
GDP (Billions NCU)		South Korea														
	Year	1995	1996	1997	1998	1999	2000	...	2009							
	OECD	409,654	460,953	506,314	501,027	549,005	603,236	...	1,065,037							
	WHO	415,773	467,645	511,990	504,659	551,983	603,236	...	1,063,059							
	WB	398,838	448,596	491,135	484,103	529,500	603,236	...	1,063,059							
	UN	409,654	460,953	506,314	501,027	549,005	603,236	...	–							
PHE % GGE		Switzerland	Luxembourg													
	Year	2007	2002	2003												
	WHO	19.54	14.77	14.02												
	WB	19.54	14.77	14.02												
	IMF	6.00	11.28	11.45												
PHE % THE		Belgium														
	Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	OECD	76.80	78.18	75.38	74.80	74.58	74.61	75.41	73.80	74.78	76.02	75.93	73.87	73.51	74.96	75.10
	WHO	68.67	70.68	68.21	67.82	67.72	67.53	68.70	67.18	70.51	71.42	72.03	72.77	67.95	66.78	68.35
	WB	68.67	70.68	68.21	67.82	67.72	67.53	68.70	67.18	70.51	71.42	72.03	72.77	67.95	66.78	68.35
		Chile														
	Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	...	2009				
	OECD	48.18	47.17	47.14	48.11	49.86	52.10	53.54	54.51	38.81	...	47.38				
	WHO	38.27	36.63	37.34	38.28	39.90	41.64	42.95	43.76	38.81	...	47.37				
	WB	48.18	47.17	47.14	48.11	49.86	52.10	53.54	54.51	38.81	...	46.79				
		Denmark														
	Year	1995	...	2003	2004	2005	2006	2007	2008	2009						
	OECD	82.52	...	84.55	84.27	84.48	84.64	84.40	84.66	85.04						
	WHO	82.52	...	79.75	79.16	79.35	79.99	80.21	80.15	80.09						
	WB	82.52	...	79.75	79.16	79.35	79.99	80.21	80.15	80.09						
		Finland														
	Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
	OECD	71.68	71.62	72.11	72.47	71.46	71.26	71.97	72.46	74.56	74.97	75.39	74.85	74.36	74.45	74.71
	WHO	82.52	82.42	82.28	81.99	82.21	82.43	82.67	82.94	79.75	79.16	79.35	79.99	80.21	80.15	80.09
	WB	72.01	71.91	72.18	71.81	71.48	71.06	71.83	72.31	68.27	69.01	69.47	70.16	70.15	70.70	72.05
		Australia														
	Year	1995	...	2003	2004	2005	2006	2007	2008							
	OECD	65.78	...	66.11	66.68	66.89	66.59	67.51	67.99							
	WHO	65.78	...	66.11	66.68	66.89	66.58	67.51	70.10							
	WB	65.78	...	64.54	64.56	64.47	64.16	65.40	65.40							

Notes: The currency unit for PHE, OOP, GDP is the current price.

Figure 1: Forecast intervals of THE in total millions international dollar from time series VAR models for OECD countries—Australia, Canada, Germany, and Iceland, 2010–2025, in different scales on the vertical axes



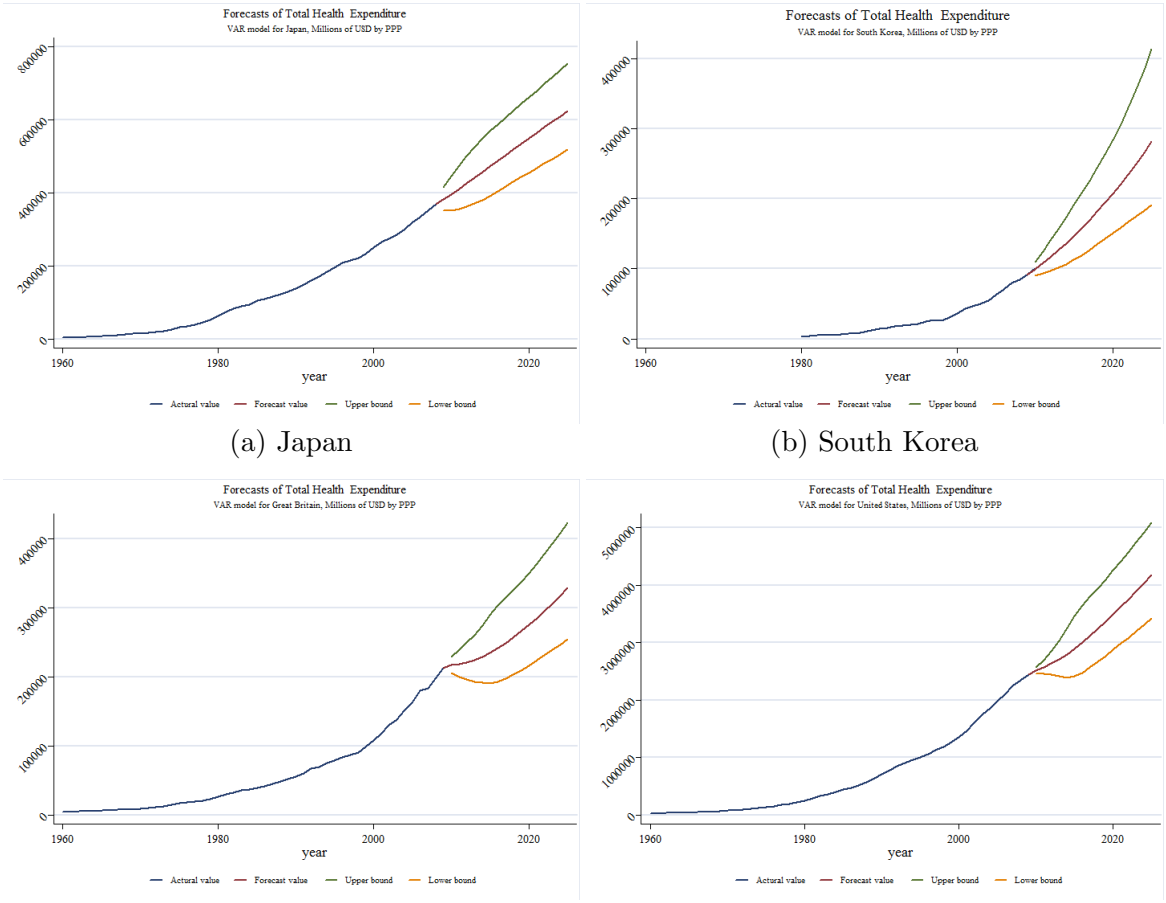
(a) Australia

(b) Canada

(c) Germany

(d) Iceland

Figure 2: Forecast intervals of THE in total millions international dollar from time series VAR models for OECD countries—Japan, South Korea, Great Britain, and United States, 2010–2025, in different scales on the vertical axes



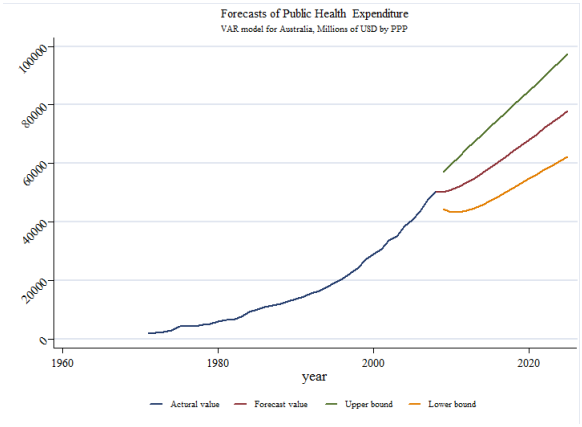
(a) Japan

(b) South Korea

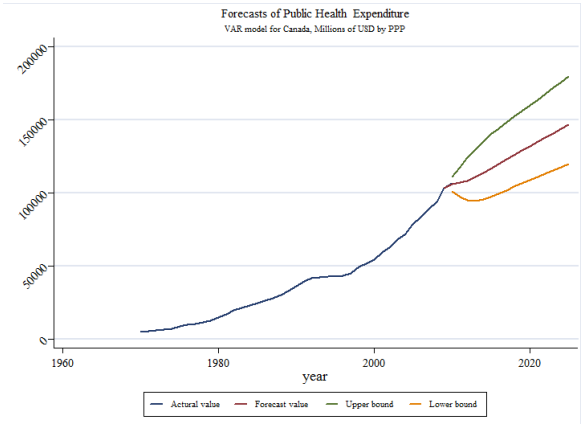
(c) Great Britain

(d) United States

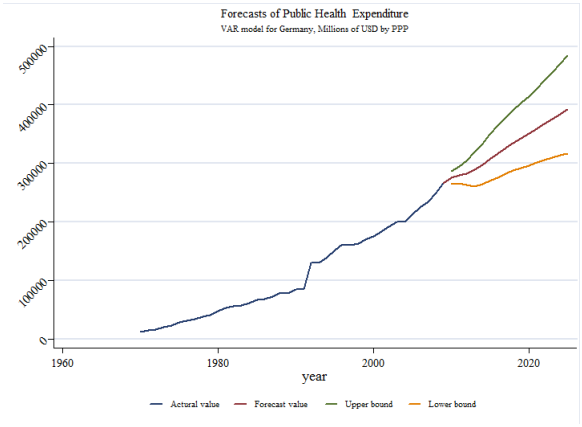
Figure 3: Forecast intervals of PHE in total millions international dollar from time series VAR models for OECD countries—Australia, Canada, Germany, and Iceland, 2010–2025, in different scales on the vertical axes



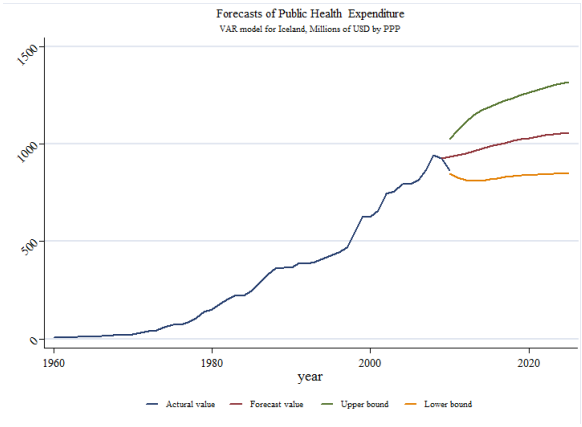
(a) Australia



(b) Canada

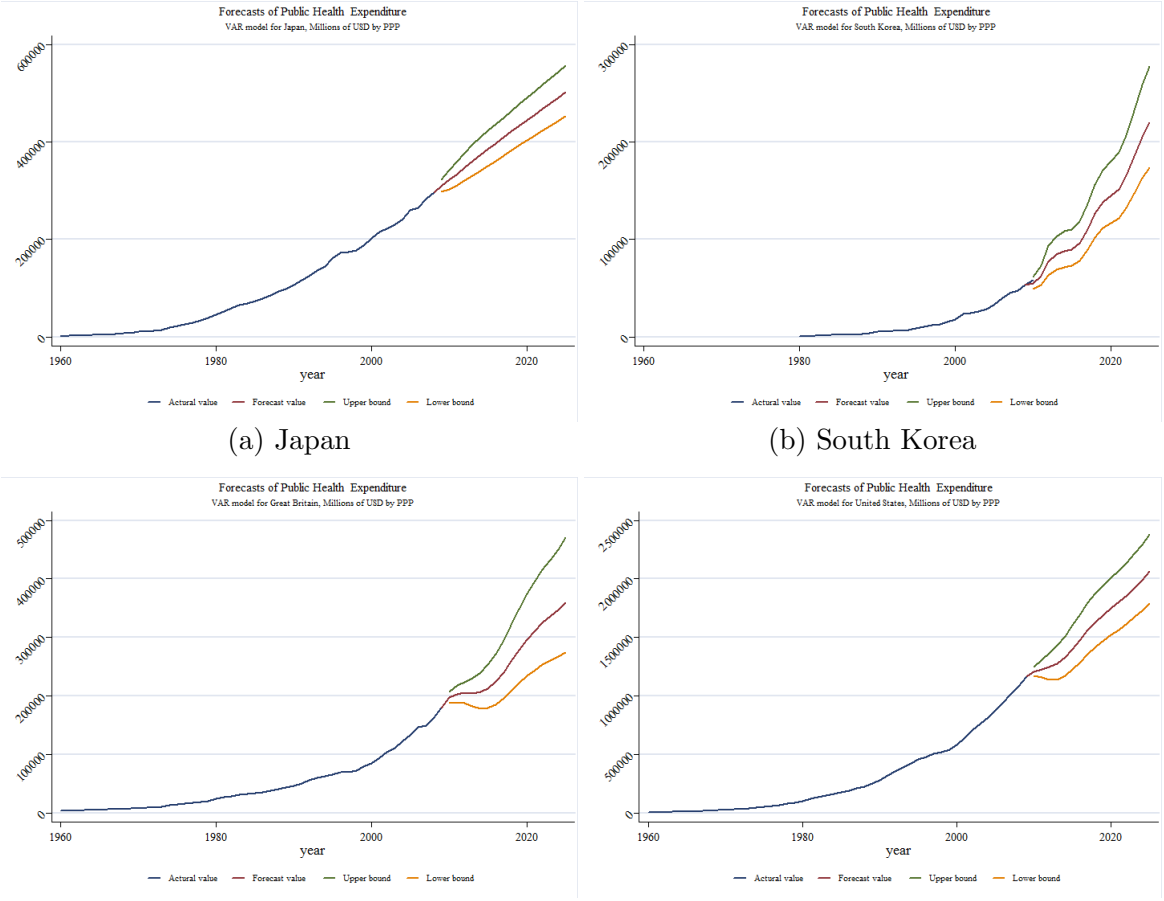


(c) Germany



(d) Iceland

Figure 4: Forecast intervals of PHE in total millions international dollar from time series VAR models for OECD countries—Japan, South Korea, Great Britain, and United States, 2010–2025, in different scales on the vertical axes



(a) Japan

(b) South Korea

(c) Great Britain

(d) United States

Bibliography

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, **58**, 277–298.

Arellano, M. and Bover, O. (1995). Another Look at Instrumental Variables Estimation of Error Components Models. *Journal of Econometrics*, **68**, 29–51.

Armstrong, J. (1978). Forecasting With Econometric Methods: Folklore Versus Facts. *Journal of Business*, **51**, 549–564.

Armstrong, J. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic.

Astolfi, R., Lorenzoni, L., and Oderkirk, J. (2012). Informing Policy Makers About Future Health Spending: A Comparative Analysis of Forecasting Methods in OECD Countries. *Health Policy*, **107**, 1–10.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley, New York.

Bank of England (2015). Inflation Report November 2005. `http://www.bankofengland.co.uk/publications/Documents/inflationreport/ir05nov.pdf`.

Barros, P. (1998). The Black-Box of Health Care Expenditure Growth Determinants. *Health Economics*, **7**, 533–544.

Blundell, R. and Bond, S. (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, **87**, 115–143.

Box, G., Jenkins, G. M., and Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken.

Cameron, A. and Trivedi, P. (2010). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Carnot, N., Koen, V., and Tissot, B. (2011). *Economic Forecasting and Policy*. Palgrave Macmillan, New York.

Centers for Medicare & Medicaid Services (1991). The Long-Term Projection Assumptions for Medicare and Aggregate National Health Expenditures.

Centers for Medicare & Medicaid Services (2000). Technical Panel Reports. Centers for Medicare & Medicaid Services.

Centers for Medicare & Medicaid Services (2004). Technical Panel reports. Centers for Medicare & Medicaid Services.

Centers for Medicare & Medicaid Services (2011). Projected Medicare Expenditures Under an Illustrative Scenario With Alternative Payment Updates to Medicare Providers.

Clemen, R. (1989). Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, **5**, 559–584.

Clements, M. and Hendry, D. (2004). *A Companion to Economic Forecasting*. Basil Blackwell, Oxford.

Clements, M. and Hendry, D. (2011). *The Oxford Handbook of Economic Forecasting*. Oxford University Press, New York.

Culyer, A. J. (1988). Health Care Expenditures in Canada: Myth and Reality; Past and Future. Canadian Tax Paper No. 82. *Canadian Tax Foundation, Toronto*.

Denton, F., Gafni, A., and Spencer, B. (2002). Exploring the Effects of Population Change on the Costs of Physician Services. *Journal of Health Economics*, **21**, 781–803.

Di Matteo, L. (2010). The Sustainability of Public Health Expenditures: Evidence From the Canadian Federation. *The European Journal of Health Economics*, **11**, 569–584.

Diebold, F. X. and Kilian, L. (2001). Measuring Predictability: Theory and Macroeconomic Applications. *Journal of Applied Econometrics*, **16**, 657–669.

Dolado, J., Jenkinson, T., and Sosvilla-Rivera, S. (1990). Cointegration and Unit Roots: A Survey. *Journal of Economic Surveys*, **4**, 249–273.

Elliott, G. and Timmermann, A. (2008). Economic Forecasting. *Journal of Economic Literature*, **46**, 3–56.

Elliott, G., Rothenberg, T., and Stock, J. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, **64**, 813–836.

European Union (2008). The 2009 Ageing Report: Underlying Assumptions and Projection Methodologies. `http://ec.europa.eu/economy_finance/publications/publication13782_en.pdf`.

European Union (2009). Long-Term Sustainability of Public Finances for a Recovering Economy. `http://ec.europa.eu/economy_finance/publications/publication15996_en.pdf`.

Evans, R., McGrail, K., Morgan, M., and Hertzman, C. (2001). APOCALYPSE NO: Population Aging and the Future of Health Care System. *Canadian Journal on Aging*, **20**, 160–191.

Fildes, R. and Makridakis, S. (1995). The Impact of Empirical Accuracy Studies on Time-Series Analysis and Forecasting. *International Statistical Review*, **63**, 289–308.

Fildes, R., Hibon, M., Makridakis, S., and Meade, N. (1998). Generalising About Univariate Forecasting Methods: Further Empirical Evidence. *International Journal Of Forecasting*, **14**, 339–358.

Gerdtham, U.-G. and Jonsson, B. (2000). International Comparison of Health Expenditure. In A. Culyer and J. P. Newhouse, editors, *Handbook of Health Economics*. Elsevier.

Gerdtham, U.-G., Sogaard, J., Jonsson, B., and Andersson, F. (1992). A Pooled Crossed Section Analysis of the Health Expenditure of the OECD Countries. In P. Zweifel and H. Frech, editors, *Health Economics Worldwide*. Kluwer Academic Publishers, Dordrecht.

Getzen, T. (2000). Forecasting Health Expenditures: Short, Medium, and Long Term. *Journal of Health Care Finance*, **26**, 56–72.

Getzen, T. (2006). Aggregation and the Measurement of Health Care Costs. *Health Service Research*, **41**, 1938–1954.

Getzen, T. (2007). Modelling Long Term Healthcare Cost Trends. *The Society of Actuaries*.

Getzen, T. and Poullier, J.-P. (1992). International Health Spending Forecasts: Concepts and Evaluation. *Social Science and Medicine*, **34**, 1057–1068.

Granger, C. (1989). *Forecasting in Business and Economics*. Academic Press, San Diego.

Harris, R. and Tzavalis, E. (1999). Inference for Unit Roots in Dynamic Panels Where the Time Dimension is Fixed. *Journal of Econometrics*, **91**, 201–226.

Heckman, J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics*, **115**, 45–97.

Hendry, D. and Hubrich, K. (2012). Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate. *Journal of Business and Economic Statistics*, **29**, 216–227.

Holden, K., Peel, D., and Thompson, J. (1990). *Economic Forecasting: An Introduction*. Cambridge University Press, Cambridge.

Im, K., Pesaran, M., and Shi, Y. (2003). Testing for Unit Roots in Heterogeneous Panels. *Journal of Econometrics*, **115**, 53–74.

Lago-Peñas, S., Cantarero-Prieto, D., and Blázquez-Fernández, C. (2013). On the Relationship Between GDP and Health Care Expenditure: A New Look. *Economic Modelling*, **32**, 124–129.

Levin, L., Lin, C.-F., and Chu, S. (2002). Unit Root Tests In Panel Data: Asymptotic and Finite-Sample Properties. *Journal of Econometrics*, **108**, 1–24.

Lin, K.-P. (2012). Time Series Analysis and Forecasting. `http://web.pdx.edu/~crkl/ec572/ec572-1.htm`.

Lorenzoni, L., Belloni, A., and Sassi, F. (2014). Health-Care Expenditure and Health Policy in the USA Versus Other High-Spending OECD Countries. *The Lancet*, **384**, 83–92.

Makridakis, S. and Winkler, R. (1983). Average of Forecasts: Some Empirical Results. *Management Science*, **29**, 987–996.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., and Simmons, L. (1993). The M2-Competition: A Real Time Judgmentally Based Forecasting Study (With Comments). *International Journal of Forecasting*, **9**, 5–30.

Makridakis, S., Wheelwright, S., and Hyndman, R. (1998). *Forecasting: Methods and Applications*. John Wiley & Sons, Hoboken.

McNees, S. (1986). Forecasting Accuracy of Alternative Techniques: A Comparison of U.S. Macroeconomic Forecasts. *Journal of Business and Economic Statistics*, **4**, 5–15.

Narayan, P. and Popp, S. (2012). A Nonlinear Approach to Testing the Unit Root Null Hypothesis: An Application to International Health Expenditures. *Applied Economics*, **44**, 163–175.

NBER (1966). *Long-Rang Economic Projection: A Report of the National Bureau of Economic Research. Conference on Research in Income and Wealth*. Princeton University Press, Princeton.

Newhouse, J. (1977). Medical Care Expenditures: A Cross-National Survey. *Journal of Human Resources*, **12**, 115–152.

OECD (2003). Projecting OECD Health and Long-Term Care Expenditures: What Are the Main Drivers? OECD Economics Department Working Papers, No. 477. *Paris and Washington, D.C.*

OECD (2011). A System of Health Accounts. `http://www.oecd-ilibrary.`

org/social-issues-migration-health/a-system-of-health-accounts_9789264116016-en.

Panopoulou, E. and Pantelidis, T. (2011). Convergence in Per Capita Health Expenditures and Health Outcomes in the OECD Countries. *Applied Economics*, **44**, 3909–3920.

Parkin, D., McGuire, A., and Yule, B. (1987). Aggregate Health Expenditures and National Income: Is Health Care a Luxury Good? *Journal of Health Economics*, **6**, 109–127.

Parkin, D., McGuire, A., and Yule, B. (1989). What Do International Comparisons of Health Expenditures Really Show? *Community Medicine*, **11**, 116–123.

Philips, P. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*, **75**, 335–346.

Samuelson, P. A. (1980). *Economics: An Introductory Analysis*. McGraw-Hill, New York.

Sims, C. (1972). Money, Income, and Causality. *The American Economic Review*, **62**, 540–552.

Theil, H. (1966). *Applied Economic Forecasting*. North-Holland, Amsterdam.

Westerlund, J. (2007). Testing for Error Correction in Panel Data. *Oxford Bulletin of Economics and Statistics*, **69**, 709–748.