Department of Finance
Ministère des Finances

Working Paper
Document de travail

Health Spending, Health Outcomes, and Per Capita Income in Canada: A Dynamic Analysis

Kathleen Day *
kmday@uottawa.ca

Julie Tousignant **
Tousignant.Julie@fin.gc.ca

Working Paper 2005-07

June 2005

* Department of Economics, University of Ottawa, 200 Wilbrod Street, Ottawa, K1N 6N5.
** Economic and Fiscal Policy Branch, Department of Finance, 140 O'Connor Street, Ottawa, K1A 0G5.
The authors would like to thank Isabelle Amano, Alison McDermott, Chris Matier, Gabriel Rodriguez, Junsoo Lee and Mark C. Strazicich for providing comments and suggestions as well as software support and useful documentation.

Abstract

While there has been much discussion of the rising cost of the health system in Canada, there has been relatively little analysis of the relationship between spending on health and population health status in Canada. This study builds on the existing literature by attempting to estimate a dynamic model of the relationship between three variables: real per capita GDP, real per capita spending on health and an indicator of health outcomes. Unit root and cointegration tests, with and without allowances for structural break(s), are used to help identify the appropriate dynamic model. Generalized impulse response analysis is then used to explore the dynamic relationships between the three variables. Several different indicators of health outcomes are employed in the analysis: the infant mortality rate, the age-standardized mortality rate, and a single composite index. The latter is constructed by applying principal components analysis to a set of common health indicators, resulting in an index that captures a high proportion of the total variation in the different indicators. Results for Canada will be compared to those of other OECD countries when possible.

The analysis presented in this paper finds evidence of a weak statistically significant relationship between per capita health spending, health outcomes, and per capita GDP. The absence of a strong statistical relationship may be due to model misspecification or may reflect the fact that at high levels of population health, the returns to increases in health spending are small.

Résumé

Bien qu'il y ait beaucoup de discussions entourant la croissance des coûts du système de santé au Canada, il existe peu d'analyses étudiant la relation entre les dépenses de santé et l'état de santé de la population au Canada. Ce document de recherche s'appuie sur la littérature existante et tente d'élaborer un modèle dynamique de la relation entre trois variables : le PIB réel par habitant, les dépenses de santé réelles par habitant et un indicateur de l'état de santé de la population. Les tests de racine unitaire et de co-intégration, avec ou sans bris structurel(s), permettent d'identifier le modèle dynamique approprié. L'analyse générale par sentiers de réponse est par la suite utilisée pour explorer la relation dynamique entre les trois variables. Plusieurs indicateurs de l'état de santé de la population sont utilisés lors de l'analyse : le taux de mortalité infantile, le taux comparatif de mortalité, ainsi qu'un indice composé. Ce dernier est calculé à l'aide de la méthode des composantes principales qui est appliquée à un ensemble d'indicateurs de santé semblables. Cela crée un indice réunissant une portion élevée de la variation totale de ces différents indicateurs. Les résultats obtenus pour le Canada sont comparés, lorsque cela s'avère possible, avec ceux de certains pays de l'OCDE.

L'analyse présentée dans ce papier révèle l'existence d'une faible relation statistiquement significative entre les dépenses de santé par habitant, l'état de santé de la population et le PIB par habitant. L'absence d'une forte relation statistique est peut-être le résultat d'une mauvaise spécification du modèle ou indique qu'à des niveaux élevés de santé de la population, les gains provenant d'une augmentation des dépenses de santé sont faibles.

1. Introduction

While the rising cost of the health care system in Canada has been a hot topic of discussion, relatively little attention has been placed on the relationship between spending on health and health outcomes (i.e. population health status) in Canada. This is surprising, since along with ever-increasing health expenditures comes the need to evaluate their effectiveness. Did cutbacks in per capita spending on health care in the 1990s affect the health of Canadians in any way? Are increases in spending needed to maintain the current level of health of the Canadian population? These are questions that can only be answered by studying the relationship between health status and health expenditures in Canada.

A related question is what determines health expenditures in Canada. Between 1976 and 2001, real per capita public spending on health (age-adjusted) grew from approximately $9,000 to $12,000, while real per capita private spending on health (age-adjusted) increased from $3,000 to $5,000.[1] As a proportion of GDP, total health spending increased from 7% to 9.6 % between 1976 and 2001, which implies that per capita total health spending in fact increased faster than per capita GDP during this period.[2] According to basic economic theory, if everything else is held constant and if health care is a normal good, an increase in per capita income will lead to increases in the demand for health care. However, during the same time period there have also been improvements in measures of population health status such as life expectancy. Nonetheless, despite a healthier Canadian population as compared to 1976, real per capita total health expenditures during the period have increased substantially and very few studies have attempted to explain this trend.

In this paper, the relationship between health status, health outcomes and per capita GDP in Canada is examined. Although other factors such as demographic structure, tobacco and alcohol consumption, and environmental quality also have an influence on health status, this analysis is restricted to a three-variable system (health outcomes, per capita health spending and per capita GDP) because many sources, including the World Health Organization and Health Canada, suggest that health care spending and income are important determinants of health status. In addition, it is relatively easy to obtain long time series of data on these three variables, but less easy to obtain the long time series for the other variables that would be required for the estimation of a more complex model.

This study will therefore attempt to explore the dynamic relationship between three variables: an indicator of health outcomes, real per capita income (as measured by GDP or GNP), and real per capita spending on health, using time series data for Canada. With data from the *Historical Statistics of Canada* this relationship is investigated for the longest time period possible. Testing is done with three different health indicators: the infant mortality rate, the age-standardized mortality rate, and a composite of a group of standard health indicators constructed using principal components analysis. A dynamic modeling approach is used as this type of approach provides insight into both the short- and long-run impacts of shocks to one variable on the other variables in the model, e.g. how an increase in per capita health spending is likely to affect health

[1] These figures are based on Finance Canada estimates. See Jackson and McDermott (2004).
[2] Data from CIHI (2004), Table A.1.

status in the short run and in the long run. The modeling approach used throughout the analysis assumes that all variables can influence each other.

The remainder of this paper is divided into five sections. Section 2 reviews the literature on the relationship between health status, health spending, and income and summarizes the different approaches used in modeling this relationship. Section 3 presents the econometric methods used in the paper. Unit root and cointegration tests are used to help identify the appropriate dynamic model. This econometric analysis differs from that of previous studies in that it uses relatively new unit root tests proposed by Elliott, Rothenberg, and Stock (1996), Perron and Rodriguez (2003) and Lee and Strazicich (2003,2004), as well as a new lag selection criterion for unit root tests proposed by Ng and Perron (2001).

The fourth section presents the data. Section 5 presents the results of the empirical analysis for subsets of variables during the periods 1960-1997, 1950-1997 and 1926-1999. The analysis is carried out for the 1960-1997 period for purposes of comparison with earlier studies that used post-1960 panel data for OECD countries. To determine whether the differences between this study's results and those of previous studies are unique to Canadian data, the same unit root and cointegration tests are applied to data from five other OECD countries as well. The last section of the paper presents the conclusions.

The analysis suggests that although some causal relationships between a measure of the health status of the population, real per capita GDP and real per capita health expenditures are statistically significant, these relationships are not very strong. While per capita income does appear to influence health status, there is only weak evidence of a relationship between per capita health spending – public or total – and health status. This finding may be due to model misspecification or may reflect the fact that at high levels of population health, the returns to increases in health spending are small.

2. Modeling the Relationship between Income, Health Status, and Health Spending

Although nearly all studies of the determinants of health status or health outcomes include some measure of health spending as an explanatory variable, very few studies of the determinants of health spending include health status as an explanatory variable. The exclusion of some indicator of health outcomes from studies of the relationship between per capita health spending and per capita GDP is surprising given that several studies of the determinants of health outcomes have pointed to simultaneity between health outcomes and per capita health spending as a possible source of bias (Filmer and Pritchet 1999, Kee 2001). The remainder of this section presents a brief review of the literature on the determinants of health status, followed by a summary of the literature on the determinants of health spending, and a discussion of the implications for modeling the relationships between per capita income, per capita health expenditures and health outcomes using Canadian data.

2.1 The Determinants of Health Status

Most studies of the determinants of health status use what is known as a "health production function" approach. As its name suggests, the health production function is a relationship

between some measure of health status and factors believed to influence health status. Following Gravelle and Backhouse (1987), the arguments of the health production function can be divided into three broad categories: the consumption of health care services; the consumption of goods, services, and activities likely to influence health status (which can also be thought of as lifestyle variables); and various environmental variables such as measures of air and water quality. In most empirical studies, a variety of socioeconomic characteristics are included as well, on the grounds that they are related to determinants of health status that cannot otherwise be measured.

In specifying a health production function, empirical researchers face a number of challenges. First, they must decide how to measure health status. At the individual level, health outcomes can be measured by an indicator of disability (Berger and Leigh 1989) or by self-reported health status (Rivera 2001, Deussing 2003). At the aggregate level, life expectancy, the infant mortality rate, and potential years of life lost have all been used as measures of population health.[3]

Second, they must obtain data on the various inputs into the production of health. Many studies have used per capita health expenditures as their measure of health care services consumed (e.g., Crémieux et al. 1999, Or 2000b, Thornton 2002), but a few have used measures of the quantity of inputs into the production of health care such as the number of physicians per 1,000 population (Or 2000a). Studies that restrict their attention to developed countries often include measures of the consumption of alcohol and tobacco (Crémieux et al. 1999; Or 2000a,b; Thornton 2002), since both are well-known to have potentially detrimental effects on health. Crémieux et al. (1999) and Thornton (2002), among others, also included a variable to reflect the level of education of the population to test Grossman's (1972) hypothesis that increases in education increase the efficiency with which health outcomes are produced. In addition, Crémieux et al. (1999) and Or (2000b) included measures of the consumption of certain nutrients – fat, sugar, and/or meat – that are believed to have undesirable health consequences if consumed in high quantities. Due to data limitations, only a handful of studies have been able to include direct measures of environmental quality – for example, in their international cross-section study Filmer and Pritchett (1999) included a variable measuring access to safe water, while in her studies of OECD countries Or (2000a,b) included NO_X emissions per capita as a measure of air quality. [4]

Nearly all studies include per capita income as an explanatory variable, although it does not fall into any of Gravelle and Backhouse's (1987) three categories of inputs into health production. The justification for including this variable varies across studies; in an early study Auster, Leveson, and Sarachek (1969, 414) noted that "this variable acts as a proxy for a host of other factors for which we would prefer specific measures." Other authors have focused on psychosocial factors, including the distribution of income (Preston 1975). Consequently, some authors' predictions regarding the sign of the coefficient of income are ambiguous. Auster et al. (1969), Or (2000b), and Thornton (2002) all note that while the higher living standards permitted by higher incomes may be conducive to better health, higher incomes may also be associated with more stressful lifestyles, which would tend to reduce health.

[3] These aggregate measures are discussed in more detail in section 4 of the paper.
[4] NO_X stands for nitrogen oxide.

Additional explanatory variables have also been included in various studies. However, despite improvements in data availability over the years, Gravelle and Backhouse's (1987) argument that most studies of aggregate population health suffer from omitted variable bias is likely still true. Data on measures of environmental quality and lifestyle variables in particular are difficult to obtain, even for a developed country such as Canada. In their most recent study using provincial-level data, Crémieux et al. (2005a,b) used real per capita spending on food and non-alcoholic beverages as a proxy for nutrition, because they are unable to obtain measures of the consumption of specific nutrients that are more directly related to health status.

The third challenge in specifying a health production function is to choose the functional form. The most popular function forms are either linear or linear in natural logarithms. Some studies, such as Crémieux et al. (1999), reported results for both specifications. While for most explanatory variables the coefficient estimates led to similar conclusions regardless of functional form, in a few cases the statistical significance of the estimates was affected by the choice of functional form. Gravelle and Backhouse (1987) suggest that interactions between some of the explanatory variables in health production functions should be considered.

In addition to these specification issues, researchers who choose to estimate health production functions are likely to run into a number of additional problems, most of which are highlighted by Gravelle and Backhouse (1987). First, studies that use international data face the problem of obtaining data that are comparable across countries. The OECD's efforts to build a health database containing data that are comparable across countries are laudable, but even the OECD's *Health Data* database has changed considerably over the years. This means that it might be difficult to replicate the results of studies such as those of Or (2000a,b), which were based on data for twenty-one OECD countries over the periods 1970-1992 (Or 2000b) and 1970-1995 (Or 2000a). Using data for a single country, as do Crémieux et al. (1999; 2005a,b) and Thornton (2002), can help to alleviate this problem.

Second, few empirical studies take into account the fact that the health production function is likely to be dynamic. For example, although the link between smoking and cancer is fairly well documented, it is also known that it takes many years for most smoking-related cancers to develop. Hence the current level of health status (for both individuals and the population as a whole) depends not only on current smoking behaviour, but also on smoking behaviour in the past. Including lagged health status as an explanatory variable in the health status equation would be a relatively parsimonious way of dealing with these dynamic relationships, but Lichtenberg (2004) appears to be the only published study that actually does so.[5] Lichtenberg finds the coefficient of this variable to be highly significant in all of the equations he estimates.

Third, simultaneity bias is a potential problem in the estimation of health production functions, because lifestyle variables, health expenditures, and income are likely to be determined simultaneously with health status. In Grossman`s (1972) human capital model of the demand for health care, individuals chose expenditures on health care and other goods subject to a health production function as well as a budget constraint; thus consumption of health care and other goods and the level of the stock of health capital are indeed simultaneously determined in his

[5] Or (2000a, 20) notes that in previous work she has found that adding lagged values of the explanatory variables in her equations did not change the qualitative results.

model.[6] Similarly, several recent studies of economic growth treat per capita income as a function of the stock of health capital, typically measured by life expectancy, implying that per capita income is likely to be correlated with the error term in a health production function.[7] To deal with this problem, a number of recent studies of the health production function, such as Kee (2001) and Thornton (2002), have used an instrumental variables (IV) estimation technique. However, finding instrumental variables that are truly exogenous is difficult. For example, is it appropriate to use, as did Thornton (2002), variables such as the per capita consumption of alcohol and cigarettes as instruments? Gravelle and Backhouse (1987) argued that since the levels of consumption of these goods and of health spending can be viewed as the outcome of the same individual utility maximization problem, they should be treated as endogenous rather than exogenous.

Finally, studies using aggregate time series or panel data face the problem that measures of aggregate population health and many of the explanatory variables included in health production functions are clearly nonstationary. For example, in the Canadian case, since 1926 life expectancy has been increasing, while the infant mortality rate has been decreasing (Figure 2). Hence the results of studies that do not take these trends into account could be spurious, as they might reflect the common trends shared by the variables rather than a real behavioural relationship.

Although the spurious regression problem has received much attention in the literature on the determinants of aggregate health spending, it does not seem to have received equal attention in the literature on health production functions. Lichtenberg (2004) appears to be the only study that addressed this issue by testing variables for unit roots. Despite evidence that some of his variables may contain unit roots, he chose not to modify his estimating equations to take them into account, but cautions the reader that "the danger of spurious correlation cannot be entirely ruled out" (Lichtenberg 2004, 383).[8]

Given all of these potential problems, it is not surprising that the literature contains many conflicting results regarding the relationship between health status and per capita health expenditures. Some recent studies have found a positive relationship between spending on health and health outcomes (Or 2000a,b; Baldacci et al. 2002; Berger and Messer 2002), but others did not find a significant relationship between the two variables (Filmer and Pritchett 1999, Thornton 2002). Still others, such as Baldacci et al. (2002), found that their results depend on the data set and/or estimation methods used. All studies, however, did find a positive and significant relationship between health outcomes and real per capita income.

To date, only a handful of papers on this topic appear to have focused on Canada, although Canada was included in the panel data sets used by Or (2000a,b) and Hitiris and Posnett (1992).

---

[6] In Grossman`s model, health is viewed as a type of capital and thus the "health production function" is really a net investment equation.

[7] See, for example, Knowles and Owen (1995) or McDonald and Roberts (2002).

[8] Lichtenberg justified his decision to proceed as if unit roots were not a problem on the grounds that (i) unit root tests are known to have less power to reject the null of a unit root when the sample period is short; (ii) the estimated value of the autoregressive parameter was not very close to 1; and (iii) he did not find a unit root in his dependent variable over the longer time period of 1900-2001. It should also be noted that he uses only the Phillips-Perron unit root test, rather than more recent unit root tests, which have been shown to be more powerful.

Hanratty (1996) used county-level data for the period 1960-1975 to examine the impact of the introduction of universal health insurance programs in Canada on the infant mortality rate. However, she did not include per capita health spending in her equations; instead, she focused on the coefficient of a variable indicating when public health insurance was introduced and the extent of coverage. She concluded that the introduction of public health insurance did significantly reduce infant mortality rates.

Crémieux et al. (1999) examined the relationship between health indicators such as infant mortality rates and life expectancy and total (public and private) per capita spending on health, using pooled time-series cross-section data for the ten provinces for the period 1978-1992. Crémieux et al. (2005a,b) estimated a similar model using data for the period 1981-1998, but disaggregated per capita health spending into three categories: public spending on drugs, private spending on drugs, and non-drug health care spending. Kee (2001), in an unpublished M.A. thesis, also used pooled time-series cross-section data for the ten provinces for the 1975-1996 period. Similar to Crémieux et al. (1999), Kee regressed indicators of population health status (infant mortality rates, life expectancy, and age-standardized mortality rates) on a number of variables, including real per capita public expenditures on health. However, unlike Crémieux et al. (1999), who used a pooled generalized least squares estimation procedure, Kee used instrumental variables estimation to control for possible simultaneity between health status and public spending on health. All three of these studies found a statistically significant relationship between health status and both health spending and per capita income. However, none of these studies estimated dynamic models.[9] Finally, Deussing (2003), in another unpublished M.A. research paper, used microdata from the 1996 National Population Health Survey for Canada, and found that provincial government spending on health does not have a statistically significant impact on self-assessed health status.

2.2 The Determinants of Health Spending

While there are many empirical papers examining the determinants of real per capita health expenditures, as noted by Gerdtham and Jönsson (2000) one of the limitations of this literature is its lack of theoretical foundations. Although at the individual level theoretical models of the demand for health care exist (e.g., Grossman 1972), empirical models of per capita health care spending at the aggregate level are somewhat ad hoc. Virtually all models assume that per capita health expenditures depend on per capita income or GDP. Demographic variables such as the proportion of the population aged 65 and over, and variables reflecting the characteristics of the health care system are also commonly included. For example, in their study of health care spending in nineteen OECD countries, Gerdtham et al. (1992) included the relative price of medical care, the number of practicing physicians per 1,000 population, the share of total health spending devoted to inpatient care, the public sector's share of total health spending, a dummy

[9] Crémieux et al. (1999) and Crémieux et al. (2005a,b) allowed for a limited type of dynamics in the form of a correction for AR(1) autocorrelation when estimating their models. Crémieux et al. (2005a, 111) argued that their correction for AR(1) autocorrelation was asymptotically equivalent to an equation in first differences and hence is equivalent to allowing for a unit root. However, the length of their time series is short (eighteen years or less, depending on the equation), so even if a first-difference model were appropriate substantial bias in the parameter estimates could remain.

variable indicating whether outpatient care is fee-for-service, another dummy variable indicating whether or not hospitals are subject to global budgeting, the female labour force participation rate, and a measure of urbanization, in addition to per capita GDP and a measure of the population age structure . Like those of many other studies, their estimating equations were log-linear in form.

A number of studies of this type based on international cross-section data, including Hitiris and Posnett (1992) and Gerdtham et al. (1992), found the income elasticity of per capita health expenditures was greater than one, implying that health care is a luxury good[10] rather than a necessity. This finding was somewhat disturbing to those who felt that basic health care should be considered a necessity, and led to a search for possible explanations. Hansen and King (1996) were the first to suggest that these earlier results might be spurious, due to nonstationarities in the data on per capita health spending and per capita GDP; to support their argument, they applied ADF tests for unit roots to data for twenty OECD countries for the period 1960-1987. After concluding that most series were difference stationary, they used the Engle-Granger test to test for cointegration, and found almost no evidence of cointegration. They concluded that the results of previous studies that did not take into account the dynamic behaviour of the data series were not reliable.

Subsequently, many other papers have tested for cointegration between per capita health expenditures and per capita GDP, sometimes including other variables in the analysis as well. Blomqvist and Carter (1997) also used OECD data, for a sample of twenty-four OECD countries over the period 1960-1991, but focused on the relationship between just two variables, per capita health expenditures and per capita GDP.[11] After testing for unit roots and cointegration, they found some evidence that the income elasticity of health care spending was in fact less than one, although they also cautioned that there was evidence that it varied considerably across countries. Since the publication of their paper, innovations in this literature have mainly consisted of applying different unit root and cointegration tests to panel data for OECD countries (McCoskey and Selden 1998; Hansen and King 1998; Gerdtham and Löthgren 2000, 2002; Okunade and Karakus 2001; Jewell et al. 2003). However, with the exception of Hitiris and Posnett (1992), all these studies have excluded a measure of health outcomes from their analysis. Furthermore, they focused solely on the long-run relationship between the variables and do not explore the short-run dynamics. In contrast, Roberts (1999) built a dynamic model of the determinants of health spending and also drew attention to the issues of parameter heterogeneity across countries and the sensitivity of parameter estimates to the presence or absence of a time trend in her dynamic model. Her results suggested that not only was it important to pay attention to the dynamic specification of models, but that further research focusing on individual countries could be worthwhile.

To date, although Canada is included in all studies that use OECD data, there appears to have been only a small number of empirical studies of the determinants of health spending in Canada. One example is Di Matteo and Di Matteo (1998) that used data for the Canadian provinces for the period 1965-1991 to examine the determinants of provincial government health spending.

[10] A superior, or luxury good, is defined as a good for which income elasticity of demand is greater than 1; in other words, demand for the good increases more than proportionately with increases in income.
[11] They also include a time trend in the relationship.

Using a log-linear specification, they regressed real per capita provincial government health spending on real per capita GDP, the proportion of the population over age 65, real per capita federal transfers to the provinces, a series of provincial dummy variables, and two dummy variables related to the provincial effects of Established Program Financing for health and post-secondary education. Like Crémieux et al. (1999) and Crémieux et al. (2005a,b), they pooled their data across provinces and used a Generalized Least Squares (GLS) estimation technique correcting for both autocorrelation and heteroskedasticity. They found provincial government spending on health to be inelastic to changes in income. Furthermore, the elasticity of provincial government health expenditures with respect to income was greater than that with respect to federal transfers (0.77 versus 0.48), although both elasticity estimates were highly significant. However, like many previous authors, they paid no attention to the problem of trends in the data or possible dynamics in the relationship between health spending and income.

2.3 Implications for Modeling

Several conclusions can be drawn from this overview of the literature on the relationship between health outcomes, health spending, and income at the aggregate level. First of all, as a whole, the literature suggests that health status, per capita income, and per capita health expenditures are closely linked, although in the absence of a formal theoretical model the exact nature of the relationship between them remains unclear. Second, all three variables are likely to be jointly determined. If population health status can be viewed as a form of human capital, then income is clearly dependent on health status; at the same time, determinants of health status such as health spending and consumption of alcohol and tobacco are likely dependent on income, resulting in a relationship between income and health status. Similarly, health spending should depend on health status as well as income, since a healthier population should not need to consume as much health care. To the extent that health spending increases health outcomes, it will also indirectly affect income if healthier populations also produce higher levels of income. Thus the joint endogeneity of the variables needs to be taken into account in an empirical analysis of the relationships between them.

Third, any model of the relationship between health status, health spending, and income should take into account the dynamic behaviour of the variables. If the nonstationarity of the variables is ignored, regression results will not be reliable. Hence a dynamic approach to modeling the relationship between the three variables would be appropriate.

Finally, since some previous studies have suggested that pooling data across countries may be inappropriate, an empirical analysis for Canada alone is likely to be worthwhile. Using data for just one country will reduce the problems of data inconsistency that are inherent in studies that use international data. However, a dynamic analysis for a single country will only be feasible if sufficiently long time series of data are available. For Canada, it is possible to obtain annual time series of thirty-eight years or more for per capita GDP (or GNP), various measures of health status, and per capita health spending. Unfortunately time series of similar length are simply not available for most other variables that many other studies have included in their models. However, focusing only on the relationship between per capita GDP, per capita health expenditures and health status remains justified, as these variables are among the main

determinants of each other. Therefore, this paper focuses on building a dynamic model of the relationship between population health status, per capita health spending, and per capita income.

3. Econometric Methods

While a simple dynamic model of the relationship between real per capita health expenditures, real GDP per capita, and a measure of the health status of the population would be of interest, the dynamic properties of the individual variables need to be explored before such a model can be developed. Since the specification for a dynamic model depends on whether or not the included variables have unit roots, the analysis begins by testing for the presence of unit roots in the data during each of the time periods considered.

Many different unit root tests have been proposed and it is well known that many of them suffer from a lack of power.[12] Previous studies of the relationship between per capita health expenditures and per capita GDP have used a variety of different tests, including the augmented Dickey-Fuller (ADF) test (Hansen and King 1996, Okunade and Karakus 2001, Gerdtham and Löthgren 2000, 2002), the Phillips-Perron (PP) test (Blomqvist and Carter 1997, Okunade and Karakus 2001), the Im, Pesaran, and Shin (IPS) (2003) panel unit root test (McCoskey and Selden 1998, Gerdtham and Löthgren 2002; Okunade and Karakus 2001), and the Im, Lee, and Tieslau (ILT) (2005) panel unit root test that allows for an unknown structural break (Jewell et al. 2003). Because this study focuses primarily on a single country, panel unit root tests are out of the question. Instead, a variety of new unit root tests that have been shown to have increased power relative to previous tests are employed.

First, the ADF test with GLS detrending of Elliott, Rothenberg, and Stock (1996), known as the ADF-GLS test, is applied. Elliott et al. (1996) show that GLS detrending considerably improves the power of the ADF test in the case of series with a constant mean or a linear deterministic trend. They provide asymptotic critical values for the case of a linear deterministic trend, as well as critical values for 50, 100, and 200 observations. For the case of a constant term only, the critical values for the test are identical to those for the standard ADF test with neither constant nor trend. The standard ADF test should be used in cases where the series is believed to have a zero mean (in which case no deterministic terms are necessary).

Second, the MZ_α test, one of several modified versions of the Phillips-Perron (PP) test proposed by Perron and Ng (1996), is employed. Perron and Ng show that these modified tests are more powerful than both the PP test and the original ADF test. As recommended in Ng and Perron (2001), the test is carried out using GLS detrending. When neither a constant nor a trend is included in the test equation, the standard Phillips-Perron test is used.

In carrying out both the ADF-GLS test and the MZ_α test, it is necessary to choose the number of lagged terms to be included in the equation. Ng and Perron (2001) have demonstrated that additional gains in the power of both the ADF-GLS test and their modified Phillips-Perron tests can be achieved by using a modified version of the Akaike Information Criterion (AIC), which

[12] See Maddala and Kim (1998) for a recent survey of unit root tests and their properties.

they call the MAIC. This criterion is therefore used to choose the lag length k for both tests.[13] Ng and Perron find that in general, the modified Phillips-Perron tests with GLS detrending have better size than the ADF-GLS test, but lower power.[14] The size advantage of the modified tests with GLS detrending is particularly pronounced for time series with a negative MA coefficient.

Several unit root tests that allow for structural breaks at an unknown point in time are also employed. Perron and Rodriguez (2003) have extended the ADF-GLS test and the modified Phillips-Perron tests of Perron and Ng (1996) to allow for a single unknown structural break. Two cases are considered: a break in the slope of the trend function; and a break in both the intercept and the slope of the trend function. The break point is selected where the t-statistic testing for a unit root is minimized, and therefore, least favorable to the unit root hypothesis. The MAIC is used to select the lag length over the time interval [0.15T, 0.85T], where T is the sample size.[15]

The Lagrange Multiplier (LM) unit root tests developed by Lee and Strazicich (2003,2004) constitute an alternative approach to testing for unit roots in the presence of structural breaks. These tests have several advantages over other unit root tests that allow for structural breaks. First, they are more flexible in that they can test for one or two structural breaks rather than just one. Secondly, they allow for a structural break under the null hypothesis of a unit root as well as under the alternative hypothesis. Because previous tests assume that the null hypothesis is a unit root without break, rejection of the null using these tests is consistent with both a stationary series with a break and a nonstationary series with a break. Finally, Lee and Strazicich (2001) and Harvey et al. (2001), among others, have shown that ADF-type endogenous one-break unit root tests tend to select the wrong break point, leading to strong spurious rejections of the unit root hypothesis. The LM tests developed by Lee and Strazicich (2003,2004) do not suffer from these problems and thus may be more powerful.[16]

In carrying out these LM tests, the number of lagged terms to be included in the test equation was chosen in the time interval [0.1T, 0.9T] using the "general to specific" procedure employed by Lee and Strazicich (2003). Starting with an upper bound *kmax* for the number of lags k, if the last included lag is significant, $k = kmax$ is chosen; if not, k is reduced by 1 until the last lag becomes significant. If no lags are significant, k is set to 0. Once the test equation has been estimated for all possible break points, the break points and the corresponding LM test statistics are determined endogenously by minimizing the value of the LM statistics over all possible break points. Three different specifications of the test can be considered: Case A allows for a break in the intercept, Case B allows for a break in the trend slope and Case C for a shift in intercept and a change in trend slope. In this paper, only Cases A and C are considered, since Lee and Strazicich suggest that they are the most relevant for economic variables. They propose two

[13] The ADF-GLS unit root test, the Perron and Ng unit root tests, and automated lag length selection based on the MAIC have recently been implemented in EViews 4.1, which was used to carry out most of the analysis.
[14] Recall that the size of a test is the probability of a Type I error, while the power of a test is measured as one less the probability of a Type II error.
[15] A Gauss program for carrying out these tests was supplied by Gabriel Rodriguez.
[16] As of yet, no study exists that evaluates the accuracy of the Perron and Rodriguez (2003) test in choosing the break date, nor compares its performance to the LM tests of Lee and Strazicich (2003). However, Perron and Rodriguez (2003) do generate critical values assuming no break under the null hypothesis, one of the practices criticized by Lee and Strazicich.

test statistics, called LM_{ρ} and LM_{τ}. Since the results of the two tests are similar, only LM_{τ} is reported in this analysis.[17]

If unit root testing implies that all the variables are integrated of order zero, i.e. trend stationary, the relationship between the variables can be modeled as a vector autoregression (VAR). An unrestricted VAR model can be viewed as the reduced form of a dynamic simultaneous equations model. However, in contrast to simultaneous equations models, VAR models circumvent the need to make assumptions about which variables are endogenous and which are exogenous by treating all variables in the model as jointly endogenous, with each variable assumed to be dependent on lagged values of itself and on the other variables in the system. Economic theory enters into the specification of a VAR model primarily through the selection of variables to be included in the system.

To choose the lag length for the VAR models, a variety of information criteria and diagnostic tests were used. After estimating a VAR model, pairwise Granger causality tests are used to test whether causal relationships between the variables exist. More specifically, Wald tests of the null hypothesis that the coefficients of all lagged values of variable *x* can be excluded from the equation for variable *y* are carried out. If this hypothesis is rejected, then *x* is said to "cause" *y* in the sense that including lagged values of *x* in the equation for *y* improves its explanatory power.

Although they provide useful information on the nature of the relationships between variables in a dynamic system, causality tests do not provide any information about the sign of those relationships. In order to determine the sign of the relationships between the variables of the dynamic system, the dynamic impact of random disturbances to the system using impulse response functions is analyzed. An impulse response function presents the effect of a one standard deviation shock to one of the innovations on current and future values of the endogenous variables. After a positive shock to one variable the deviations in the variables are expected to converge back to zero, because in an unrestricted VAR model shocks do not have a permanent effect.

One of the limitations of VAR models is that because they are reduced-form rather than structural models, shocks to the error terms (i.e., the impulse responses) cannot be given a structural interpretation unless additional identifying restrictions are imposed on the model. Rather than imposing such identifying restrictions, in this paper generalized impulse responses as proposed by Koop, Pesaran, and Potter (1996) and Pesaran and Shin (1998) are computed. Unlike other methods of identifying impulse responses, the method used to compute generalized impulse responses does not require additional information, nor are the generalized impulse responses sensitive to the ordering of the equations of the model. Instead, they reflect the historical pattern of shocks during the sample period used to estimate the model. They are thus particularly useful for deriving stylized facts about the historical relationships between variables.

If all or some of the variables are integrated of order greater than zero, i.e. difference-stationary, then tests for cointegration can be applied to test for the existence of a long-run relationship between the variables. In cases where all variables are I(1), or integrated of order one, Johansen's trace test and maximum eigenvalue test, described in Johansen (1995), are applied. As far as we

[17] Gauss programs for these tests written by Lee and Strazicich can be found at http://www.cba.ua.edu/~jlee/gauss/ .

know, Johansen's tests have not been extended to allow for possible structural breaks in the cointegrating relationship. However, Gregory and Hansen (1996) have extended the single-equation Engle-Granger cointegration test to allow for a break in either the intercept or the intercept and trend of the cointegrating relationship at an unknown time. Thus when testing for cointegration, their tests as well as the Johansen tests are employed.

If cointegration exists, a vector error correction model (VECM) that incorporates the cointegration restrictions can be built. A VECM restricts the long-run behaviour of the endogenous variables in a VAR model to converge to their cointegrating relationships while allowing a wide range of short-run dynamics. Generalized impulse response functions can also be used to analyze the dynamic behaviour of this type of system. In contrast to an unrestricted VAR model, in a VECM, a temporary shock can have a permanent effect on the variables of the system.

4. Data

For the purposes of this study, data on three variables are needed: a measure of the population's health status, a measure of real per capita spending on health, and a measure of real per capita income. As there is no general agreement as to what is the best measure of the overall health of the population, several alternative measures are used, all of which are discussed in section 4.1. In addition, in order to obtain long time series of data on health spending and income per capita, certain assumptions had to be made about how to link overlapping data series. Further details about data sources and the construction of the data are provided in the appendix A, while the behaviour of the data over the sample period is discussed in sections 4.2 and 4.3.

4.1 Measures of Population Health Status

Health status, which refers to the level of health of an individual, group or population, is difficult to estimate and there is no universally accepted indicator that captures all the aspects of health. Different measures of health status are available but they provide only a partial perspective on the population's level of health. The most commonly used indicators are based on mortality data. These indicators capture the decrease in mortality rates and therefore provide an indication of improvement in the quantity of life, not in the quality of life. This means that even if these measures show an improvement in longevity, they are not sufficient to indicate that health status has improved. The following is a list of commonly used measures:[18]

1) *Potential Years of Life Lost* (PYLL) consists of the number of years of life "lost" when a person dies before age 70 or age 75. It provides an indirect estimate of how many deaths could potentially be avoided. The rate per 100,000 population is more useful because it takes into account the effect of the size of the population.

2) The *Infant Mortality Rate* (IMR) refers to the number of deaths per 1,000 live births. It generally reflects the level of mortality and the effectiveness of preventive care and the attention paid to maternal and child health.

[18] This list is taken from http://www.statcan.ca/english/freepub/82-221-XIE/free.htm .

3) The *Perinatal Mortality Rate* (PMR) consists of the count and rate of fetal deaths of 28 or more weeks gestation and infant deaths under one week per 1,000 total births. This indicator reflects standards of obstetric and pediatric care, as well as the effectiveness of public health initiatives.

4) The *Age-Standardized Mortality Rate* (ASM) is the number of deaths per 100,000 of total population, standardized for the age composition of the population. The use of a standard population adjusts for variations in population age distributions over time and across different geographic areas.

5) *Life Expectancy* (LE) is the number of years a person would be expected to live, starting from birth (for life expectancy at birth) or at age 65 (for life expectancy at age 65), on the basis of the mortality statistics for a given observation period. It measures quantity rather than quality of life.

6) *Probability of Survival from Birth to Age 80* is the probability of a newborn infant surviving to age 80, if subject to prevailing patterns of age-specific mortality rates. This measure is recent and not widely used. Statistics Canada provides it for the period 1986 to 1996.

Recently, other measures have been developed that aim at measuring the quality as well as the quantity aspect of life associated with health, such as *Health-Adjusted Life Expectancies (HALEs)* and *Disability-Adjusted Life Years (DALY).*[19] Essentially, these measures adjust life expectancy for quality of life mainly by using years of life without any activity limitation.

In this study, only five of the indicators will be used as measures of health status because of the absence of data for a long time period for certain measures (especially those that adjust for the quality of life), and because the selected indicators are the most widely used in the literature. These measures are the infant mortality rate (IMR), the perinatal mortality rate (PMR), the age-standardized mortality rate (ASM), life expectancy at birth (LEB), and life expectancy at 65 (LE65). The infant mortality rate is used for the analysis because it is available for the 1926-1999 time period. The IMR and the ASM are both used in the analysis of the 1950-97 and post-1960s periods.

Figure 1 shows the evolution of IMR, PMR, ASM, LEB and LE65 between 1950 and 1997. All the indicators show an improvement in health status between 1950 and 1997.[20] The greatest improvements (in percentage terms) over the 1950-1997 period have been in IMR and PMR, while LEB and LE65 improved the least. The improvements in IMR and LEB reflect

[19] See Jee and Or (1998) for a discussion of these and other alternative health indicators.

[20] The indicators have been converted to indexes with their first year equal to 100, thus any improvement in the measure reflects improvement in the population health status. IMR, ASM, PYLL and PMR would ordinarily show an improvement in the health status of the population through a decline in their levels (i.e, a decline in mortality means children are healthier), but to make them comparable to life expectancy in the figures their growth rates were multiplied by –1.

observations made in previous work which suggest that life expectancy at birth has increased primarily as a result of the reduction in infant mortality rates, since the effect on life expectancy is larger when mortality rates fall at younger ages. The percentage improvement in ASM lies between that of the other four measures over the period 1950-1997.

When a longer time horizon is considered, between 1926 and 1997, the greatest improvements (in percentage terms) are again observed for IMR and PMR, while LEB and LE65 improved the least (Figure 2). However, the improvement in LEB is greater when 1926 serves as the base year, while the evolution remained comparable for the other three variables. Lise (2000) notes that prior to the mid 1960s, the increase in life expectancy at birth was the result of more people surviving childhood and early working years and living to old age. Since the mid 1960s, while falling childhood mortality has continued to be an important factor, there have also been increased gains resulting from falling mortality rates over age 55, leading to increased years of old age. However, the total improvement in LEB since the mid 1960s has been only about half of the gain during the previous forty years.

Finally, for the 1950-1997 and 1960-1997 periods a composite indicator constructed from five of the individual measures is also used. While the health status indicators discussed above provide different perspectives on population health, they are all derived from vital statistics data on death rates and are highly correlated. A summary indicator that could account for a high proportion of the variation in the group of indicators considered may thus serve as a good overall indicator of health status.

To construct a summary indicator, principal components analysis was used.[21] This mechanical procedure produces a single indicator of health status (the first principal component) summarizing the information contained in multiple measures using a linear function that applies a different weight to each variable. The weights are derived from the eigenvectors of the correlation matrix. An indicator summarizing IMR, ASM, PMR, LEB and LE65 was computed (Figure 3). Its behaviour throughout the 1950 to 1997 period shows a clear improvement in population health status. Other indicators can be computed for different time periods and different subsets of the individual indicators.

4.2. Public/Private Health Spending in Canada

Figure 4 presents a brief portrait of the evolution of health spending in Canada. Since 1975, public health expenditures (real per capita, age-adjusted, Department of Finance calculations)[22] have fluctuated somewhat but have increased overall from approximately $9,000 in 1976 to $12,000 in 2001. Private health spending (real per capita, age-adjusted, Department of Finance calculations) over the same period increased from $3,000 to $5,000.

Between 1975 and 1991, public sector health spending grew rapidly, but between 1993 and 1997, spending decreased in real per capita terms. This reduction coincides with the period in which Established Programs Financing (EPF) and the Canada Assistance Plan (CAP) were

[21] For more information, see chapter 7 of Morrison (1967).
[22] See Jackson and McDermott (2004).

consolidated into the Canada Health and Social Transfer (CHST) and total cash payments under this program were reduced.

Since real per capita age-adjusted expenditures are only available for a relatively short period of time, total real health expenditures per capita, i.e. the sum of private and public spending, are used in the econometric analysis. This is done to capture the variation in spending between the two sectors and because the breakdown is not available historically as far as would be needed. However, for the 1960-1997 and 1950-1997 periods, real public health expenditures per capita are also used since these data are available.

Since 1945, total real health expenditures per capita have increased (Figure 5), apart from the 1993 to 1997 period where they remained relatively stable, reflecting the public sector spending period mentioned above. The relatively large increase observed in 1960 is the result of a break in the data series available in the *Historical Statistics of Canada*; prior to 1960, the total health expenditure data do not include the categories "other drugs and appliances," "other personal health care," and "other health expenses." In 1960, the first year for which data on these categories of spending are available, they accounted for about 34% of total health spending.

4.3 Real Income per Capita

Real income per capita is measured by real GDP per capita in most studies of the determinants of health spending and health outcomes. However, real GDP is not available for Canada for the entire 1926-1999 period. Therefore real GDP per capita is used only for the 1950-1997 and 1960-1997 periods. For the 1926-1999 period, real GNP per capita is used instead. Figure 6 illustrates the behaviour of the two series, in millions of 1997 dollars, over the 1926-1999 period. It can be seen that the two series behave in a very similar fashion over time, although the gap between them has widened somewhat in recent years.

5. Results

In this section, the results of the dynamic analysis are reported for the three different time periods analyzed: 1960-1997, 1950-1997, and 1926-1999. All tests were applied to the natural logs of the variables.

5.1 1960-1997

5.1.1 *Unit Root Tests*

As noted in section 2, a number of authors have recently tested for unit roots and cointegration between real per capita GDP and real per capita health expenditures in OECD countries, using various unit root and cointegration tests. The sample period covered by these studies varies, but all begin in 1960. Their results for Canada are summarized in Table 1.

A glance at Table 1 reveals substantial discrepancies between the previous studies on the order of integration of real per capita GDP and real per capita health expenditures in Canada. Even studies, which appear to use the same data (they cite the same sample period and data source), obtain different results. In some cases these differences may be due to the fact that not all studies tested explicitly for the order of integration; of the four studies that used data for the period 1960-1997, only Okunade and Karakus (2001) present the results of unit root tests for both the levels and the first differences of variables, which may help to explain why they are the only ones to conclude that for Canada, health expenditures are integrated of order two (using both the ADF and PP tests).[23] The ADF tests (but not the PP tests) carried out by Okunade and Karakus (2001) also imply that per capita GDP is integrated of order two for Canada. Jewell et al. (2003), on the other hand, using a panel unit root test that allows for the possibility of structural break, find that per capita GDP is trend-stationary (with no breaks) for Canada.

These discrepancies in the results of previous studies are important because they have important implications for the analysis of the dynamic relationship between a health status measure, real per capita GDP and real per capita health expenditures. The results of the ADF-GLS and MZ_α tests for three alternative indicators of health outcomes (the composite indicator, the age-standardized mortality rate, and the infant mortality rate), real GDP per capita (LGDP), real per capita total health expenditures (LTHE), and real per capita public health expenditures (LPHE) are presented in Table 2. To determine the order of integration of each variable, the ADF-GLS and MZ_α tests were applied to the levels, first differences, and if necessary, the second differences of the series. For all tests, the maximum lag length was set at 9. The choice of deterministic terms in the test equation was based on an examination of graphs of the levels (Figure 7) and first differences (Figure 8) of each data series. Since the graphs indicate that all the data series exhibit clear trends, for tests on the levels of all variables the deterministic terms are assumed to consist of a constant and a trend for the ADF-GLS and MZ_α tests; for the first differences, only a constant term is included for the two tests; and for the second differences, the standard ADF or Phillips-Perron test with neither constant nor trend is used. For the ADF, ADF-GLS and MZ_α tests, the MAIC was used to select the appropriate number of lagged terms to include in the test equation. For the PP test, the Newey-West rule was used to select the bandwidth (with Bartlett kernel).

For per capita real GDP (LGDP) and the age-standardized mortality rate (LASM), both tests imply that the variables are I(1) at the 1% level of significance. In addition, real per capita total health expenditures appear to be I(1) at the 10% level of significance. In contrast, both tests indicate that the other two measures of health status, the infant mortality rate (LIMR) and the composite indicator (LHS), are I(2). With respect to real per capita public health expenditures (LPHE), the tests provide conflicting results: the ADF-GLS test shows public health expenditures to be I(2) at the 1% level of significance, while the MZ_α test on the first difference of public health expenditure just as clearly implies that LPHE is I(1).

[23] Gerdtham and Löthgren (2002) note that ADF tests applied to the first differences of their data implied that per capita health expenditures were I(2) in 15 of 25 OECD countries examined, but they do not indicate which countries these were. They also found that the IPS (2003) panel unit root test implied that this variable was not I(2). Similarly, in footnote 15 of their 2000 paper they note that ADF tests applied to the first differences of the log of per capita health expenditures (with a constant but no trend included in the test equation) led to the conclusion that this variable was I(1) in all but four of the 21 countries they examined, but they do not indicate the countries for which per capita health spending was I(2).

While the finding that LHS, LIMR, and LPHE could be I(2) is unexpected, a glance at Figure 8 reveals some changes in the variability of the first differences of the series that could be the cause of this result. The change in variability is particularly pronounced in the first difference of LPHE (Figure 8f), and in fact looks like it could be due to a structural break in the first-differenced series rather than a unit root.[24] Table 3 presents the results of the Perron and Rodriguez (2003) ADF-GLS and MZ_α tests allowing for a single structural break for all six variables. When the tests are applied to the levels of the variables, structural changes in both the intercept and the slope are assumed; because Perron and Rodriguez (2003) do not include a case with only a break in the intercept, the same assumption is used when testing the first differences.[25] If the null hypothesis of a unit root is not rejected for the first difference, then the standard ADF and Phillips-Perron test results in Table 2 apply.[26]

For real per capita GDP and the three measures of health status, the results with respect to the order of integration are clear. According to both the ADF-GLS and the MZ_α tests, real per capita GDP is I(0) with a structural break in 1975. On the other hand, the three measures of health status are I(1) around a structural break. Both unit root tests reject the null hypothesis of a unit root at the 10% level of significance for the health status index. However, the date of the break differs between the two tests – 1975 for the ADF-GLS test and 1981 for the MZ_α test. For the age-standardized mortality rate, the ADF-GLS test implies that it is I(1) at the 1% level of significance with a break occurring in 1975; the MZ_α test leads to a similar conclusion, though at a lower level of significance. Finally, both tests indicate that the infant mortality rate is I(1) at the 1% level of significance for the ADF-GLS test and at the 5% level of significance for the MZ_α test, with breaks in 1983 and 1981 respectively.

In contrast, the results are less clear for real per capita public health expenditures. The ADF-GLS test indicates that LPHE is I(1) at the 10% level of significance (break in 1979), while the MZ_α test implies that LPHE is I(0) (break in 1966). For the last variable, real per capita total health expenditures (LTHE), the results indicate that the variable is I(2), which is consistent with the results in Table 2.

Finally, Table 4 presents the results of the Lee and Strazicich (2003,2004) LM unit root tests. Two different tests were done, one allowing for one structural break (LM-1) and another allowing for two breaks (LM-2). For tests performed on the levels of all variables, a constant, a trend, and a break in the trend (two breaks in trend and intercept for LM2-test) are included in the test equation; for the first differences, only a constant and a break in the intercept are included. This change in the assumptions about the deterministic terms seems reasonable since a break in the trend of the series in level should cause a break in the intercept of its first difference.

According to the LM unit root tests, almost every variable is I(0) around one or two breaks at a significance level of 10% or less. Only in the case of the health status index are the results of the two tests a little different. The LM-1 test rejects the null hypothesis of a unit root at the 10%

[24] Juselius (2005, 118) presents an example illustrating how a broken trend causes a level shift in the first difference of a series.
[25] The case with only a break in trend produces similar results to the case with breaks in both trend and intercept.
[26] Note that structural breaks are not relevant when neither constant nor trend are included.

level of significance for the level of the series, with a break in 1977. However, the LM-2 test implies that LHS is I(1) at the 1% level of significance with breaks in 1974 and 1976 – but the latter break date is not significant at the 10% level of significance.

When the break points selected by the Perron and Rodriguez tests are compared to those selected by the LM-1 test, it is interesting to note that for all series, except LHS, the LM-1 test selects a later break point. This is of particular interest because in their Monte Carlo analyses of the limitations of ADF-type tests that allow for structural break(s), Lee and Strazicich (2001) and Harvey et al. (2001) note that these tests tend to choose a break point that is earlier than the true break point. A comparison of the break points chosen by the LM-1 and LM-2 tests shows that the break point chosen by the LM-1 test always lies in the interval defined by the two break points chosen by the LM-2 test. Finally, nearly all the tests in Tables 3 and 4 seem to select at least one break at the end of the 1970s and beginning of the 1980s. A break point during this time period would seem logical for GDP and the health spending variables since it marks the beginning of the era of cutbacks in government spending. It is less clear why the health status indicators should display breaks at this point in time. The LM-2 test implies a second break around the late 1980s-early 1990s for LGDP, LASM, and LIMR, with earlier breaks for the health expenditure series.

To see whether the unit root tests used in this paper would yield similar results for other countries as well, the tests were applied to OECD data for Canada and five other OECD countries – Finland, Norway, Switzerland, the UK, and the USA – for the period 1960-1997. The results are summarized in appendix B. In general, the tests yield similar results for the order of integration. For Canada, Finland, Norway, the U.K., and the U.S., both real per capita GDP and real per capita total health spending are found to be stationary; the LM-1 test indicates that the infant mortality rate is trend stationary as well in all five countries.

Overall, the results of the unit root tests that allow for structural breaks suggest that the earlier finding that some variables were I(2) was due to a failure to take structural breaks into account. Once structural breaks in the data are allowed for, all the variables appear to be either trend stationary with a break in the trend, or I(1) with a break in trend. Because, as was discussed earlier, some ADF-type tests have been shown to have problems, in this paper it was decided that more weight would be placed on the LM tests. The results of these tests imply that there is no need to undertake cointegration testing.[27] Instead, a simple VAR model with structural breaks included would be sufficient to investigate the dynamic relationships between the variables (with the possible exception of LHS, for which the two LM tests provide different results regarding the order of integration). Thus, in the remainder of this section, the properties of some VAR models involving the variables under consideration are presented.

[27] Cointegration analysis under the assumption that the variables were I(1) was also carried out. The results of these tests and analyses of the VECM models did not produce conclusive results. It should be noted that although the Johansen tests performed did not take possible structural breaks into account, the modified Engle-Granger tests allowing for one structural break proposed by Gregory and Hansen (1996) were also carried out. The Johansen tests did not show signs of cointegration in most cases, while cointegration was present in most cases using the Gregory and Hansen tests.

5.1.2 *VAR Analysis*

Based on the results of the previous section, many different three-variable VAR models consisting of a health status indicator, real per capita GDP, and a measure of health spending could be constructed: six possible combinations of the three variables are possible, and different variations on each of these combinations can be obtained by including different dummy variables for structural breaks. In this section, only the three models involving the age-standardized mortality rate (LASM) are considered.[28] Dummy variables reflecting the break points identified by the LM unit root tests were included in the models.[29] Each dummy variable was included in all the equations of the VAR model. For all models, the lag length used was that chosen for an unrestricted VAR model with dummy variables. In section C of the appendix, the results of tests used to determine the appropriate number of lags for the unrestricted VARs are presented.

Table 5 contains Wald statistics for pairwise Granger causality tests for three VAR models involving LASM, with the p-values of the statistics in parentheses. The difference between Models 9A and 9B, both of which include LTHE, is that the first model includes the dummy variables suggested by the LM-1 test, while the second model includes the dummy variables suggested by the LM-2 test. Model 9C includes LPHE rather than LTHE, as well as the dummy variables suggested by the LM-1 test. The table shows that for all three VAR models examined, the null hypothesis that per capita GDP does not "cause" LASM at levels of significance of 4% or less can be rejected. However, the relationship between the two variables appears to be unidirectional, since the null hypothesis that LASM does not cause LGDP cannot be rejected. In other words, changes in income cause changes in health status, but income is not itself dependent on health status. This dependence of health status on income may be due to the dependence of excluded determinants of health status, such as tobacco and alcohol consumption, on income.

With respect to the relationship between LASM and per capita health expenditures, the results are somewhat sensitive to the model specification. Interestingly, the evidence in favour of causality running from health status to per capita health expenditures is stronger than the evidence in favour of causality running in the reverse direction; for two of the three VAR models the results imply that health status has a significant impact on health spending at the 5% level of significance, while for none of the models does health spending have a significant impact on LASM at this level of significance. Per capita public health expenditures do have a significant impact on LASM at the 10% level of significance in Model 9C, though.

Finally, the results in Table 5 suggest that causality between LGDP and the two alternative measures of health spending is also unidirectional at best, from LGDP to health spending. That this is the case is not surprising, since as discussed in Section 2, the effect of health spending on income is likely to be indirect, resulting from health spending's impact on health status. Since health status is also included in the model, this indirect channel is already accounted for. However, in model 9A there appears to be no relationship between the two variables at all. Since

[28] Results for additional models are available from the authors upon request. It should be noted that the analysis of impulse response functions of these models produced similar results.

[29] The dummy variables were defined to be consistent with Lee and Strazicich (2003). For example, if 1979 was identified by the LM unit root test as a break point, the dummy variable was defined to equal to zero for $t \leq 1979$ and one for $t > 1979$.

Models 9A and 9B differ only in the choice of dummy variables to reflect structural breaks, the results indicate that the choice of break points is indeed important.

Thus overall, the results of the causality tests confirm that the variables of the model are related, although not all the relationships predicted by economic theory are validated. Per capita income clearly has an important impact on health status, even after controlling for either public or total health spending per capita. There is also evidence that both public and total health spending per capita depend on income per capita, as previous research suggests. Finally, there is also some evidence of a relationship between health status (as measured by LASM) and per capita public health spending, although in terms of statistical significance, the impact of health status on health spending appears to be more important than the reverse relationship.

Although the causality tests provide some insight into which variables influence each other, they do not tell us whether the overall impacts are positive or negative. The generalized impulse response functions presented in Figure 9 show what happens when a temporary one-standard deviation shock to the error term in one equation occurs. It should be noted that a positive shock to LASM implies a deterioration of health status, rather than an improvement. As mentioned in Section 2, in a VAR model, the deviations in the variables are expected to converge back to zero after a positive shock to one variable, because the shock does not have a permanent effect.

The results in Figure 9 show that the variables adjust very slowly to shocks – after a one-standard deviation shock to a variable, it takes more than fifty years for the effect of the shock to dissipate. The same was true in models consisting of other combinations of variables. The model of Figure 9A does not even converge; however, this may be due to the strong colinearity between the dummy variables included in the model. The LM-1 test identified breaks in 1979, 1980, and 1982 for the three variables of the model, three years, which are very close together. When the VAR was re-estimated with a dummy variable for 1980 only, the impulse responses do converge. However, including dummies for the adjacent years 1979 and 1980 did not seem to cause convergence problems for Model 9C.

Looking first at the direction of the responses in Figure 9 (panels B and C), one can see that most of the results are consistent with one's prior expectations. Shocks to the health spending equation of the models decrease age-standardized mortality in the medium term (after a slight increase in the first years), before the shock dissipates. Similarly, one would expect a shock that increases age-standardized mortality to have a negative effect on GDP in the short run, and this is confirmed in Models 9B and 9C, although there is a slight increase in GDP in the early years. Shocks to GDP also seem to have the expected effect, reducing age-standardized mortality and increasing per capita total and public health spending. The impact in the first year is, though, sometimes negative.

One result that does not seem to be consistent with prior expectations is the effect on per capita total health expenditures of a shock to age-standardized mortality. If this were a structural model, one would expect a positive shock to LASM, which deteriorates health status, to increase health expenditures. Instead, in Model 9B, real per capita total health expenditures decrease following a positive shock to the age-standardized mortality rate (i.e., a decrease in the health status), and

then slowly increase in the long run as the shock dissipates. However, in Model 9C per capita public health expenditures increase as expected.

Unfortunately, although most of the responses seem to be in the right general direction, most of them are also very small – less than 1% in magnitude. Furthermore, all the generalized impulse responses shown in Figure 9 lie within their 95% asymptotic confidence bounds.[30] Thus even though the causality tests indicated that there are some statistically significant relationships between the variables, the VAR models are estimated so imprecisely that the impulse responses are not statistically significant. This means that even though the impulse responses generally show the expected impacts, the lack of statistically significant results makes it difficult to draw any economic or policy implications. However, even if the impulse responses are not statistically significant, it is interesting to note that the impact of a shock to per capita health spending does improve health status, which suggests that public investments in health spending may not be totally inefficient.

5.2 1950-1997

5.2.1 *Unit Root Tests*

For the 1950-1997 period, the results of the ADF-GLS and MZ_α tests applied to Canadian data are presented in Table 6. The levels and first differences of all variables are graphed in Figures 10 and 11 respectively. Figure 10 shows that there are trends in all variables, while an outlier in total health expenditures resulting from the inclusion of new categories of expenditure in 1960, is evident in Figure 11b. There is also some evidence of a decrease in the variability of the first differences of LHS and LPHE during the latter half of the sample period.

Since the behaviour of all the variables is similar to that observed during the 1960-1997 period, the same assumptions about constants and trends were made for the unit root tests. For all of the variables, the results of the ADF-GLS test with respect to the order of integration during the period imply that the series are I(2). The test indicates that the second differences of all variables are definitely stationary at the 1% level of significance, but neither the levels nor the first differences are stationary at even the 10% level of significance.

The results of the MZ_α test are consistent with those of the ADF-GLS test, except in the cases of the two health spending variables. The MZ_α test implies that real per capita total health spending is I(1), not I(2). For this variable, the difference between the results of the two tests may be due to the presence of the large outlier in LTHE in 1960, resulting from the previously-mentioned break in the data. According to Vogelsang (1999), the modified Phillips-Perron tests are more robust to additive outliers and therefore the MZ_α test may be more reliable in this case.

Although the MZ_α test implies that per capita *public* health spending is I(1) during the 1960-1997 period, the test implies that during the 1950-1997 period, it is I(0); more specifically, that it is trend stationary. The ADF-GLS test, on the other hand, implies that public health spending is I(2) in both periods. It is hard to explain why the results of the two tests are so different for this variable; however, in a simple ARMA(1,1) model for the first difference of per capita public

[30] The confidence bounds are not shown in the figures because of space limitations.

health spending there was a negative MA coefficient of –0.11. Whether this negative coefficient means that the MZ_α test is the most reliable of the two, for this variable, also is unclear as the negative MA coefficient is not significantly different from zero.[31]

Another possible explanation for some of the inconsistencies in the results, and the apparently high order of integration of some of the variables, may be that structural breaks have occurred. Since evidence that structural breaks occurred in the shorter 1960-1997 time period is found, structural breaks are almost certainly important during the 1950-1997 period also. The results of the Perron and Rodriguez (2003) ADF-GLS and MZ_α tests allowing for a structural break are presented in Table 7.[32] Compared to the 1960-1997 period, the results with respect to order of integration are different for all the variables, except real per capita public health expenditures. The ADF-GLS test implies that real per capita public health is I(1) with a break in 1961, while the MZ_α test implies that LPHE is I(0) with a break in 1976.

As far as the other variables are concerned, in comparison to the 1960-1997 sample period, with respect to order of integration, the two tests now provide conflicting results for real per capita GDP and age-standardized mortality. However, the extension of the sample period does not seem to increase the order of integration of any variable, except the infant mortality rate. Both tests now indicate that the infant mortality rate is I(2). All other variables are found to be I(1) by at least one test, except LHS which appears to be I(0) according to both tests. LGDP and LASM also appear to be trend stationary around a structural break according to the MZ_α test. The locations of the breaks, however, are different from those obtained using the 1960-1997 sample.

The results of the Lee and Strazicich (2003,2004) LM tests for the 1950-1997 sample period are presented in Table 8. The results of the LM-2 tests are very similar to those for the 1960-1997 sample period: all the variables, except LHS, are found to be trend stationary around two structural breaks. Using the LM-1 test, it is no longer possible to reject the null hypothesis of a unit root in the levels of LGDP and LPHE. As well, LHS is now I(2), while it was I(0) in the 1960-1997 period. This is not necessarily surprising, since the longer the sample period, the greater the likelihood that there is more than one structural break, and the higher the risk of rejecting the null hypothesis of a unit root when it is not true. Thus the differences in results between the LM-1 and LM-2 tests suggest that for this sample period the test that allows for more breaks is the more useful of the two.

Comparing the locations of the two structural breaks across the 1950-1997 (Table 8) and 1960-1997 (Table 4) samples, one can see that the locations of the breaks are different for all variables except LIMR, the infant mortality rate. For LASM, the date of the first break does not change much (1974 vs. 1975), but the date of the second break changes from 1993 to 1987. Real per capita GDP now seems to have only one statistically significant break, in 1970, rather than breaks in 1980 and 1989. It is hard to explain why there should be such big changes in the selected break dates when the sample size changes, but in both cases the sample size is not large. Sampling variability is thus likely to be a problem. It is also possible that when the sample period is lengthened, observations that may have seemed indicative of a break in the smaller sample no longer seem out of line.

[31] Its p-value is 0.691.
[32] The assumptions made about deterministic components are the same as in the previous section.

Overall, the results seem to confirm the conclusion of the previous section that structural breaks in the variables are important and need to be taken into account when testing for unit roots. Moreover, there is some evidence that a test that allows for more structural breaks is better for this extended sample period. Thus, rather than testing for cointegration, the next section will look at the behaviour of VAR models based on the results of the LM-2 test.

5.2.2 *VAR Analysis*

The results of the LM-2 tests discussed in the previous section suggest that real per capita GDP, both health spending variables, and the health outcomes indicators, LASM and LIMR, are trend stationary. Thus, as in the previous section, four trivariate models consisting of LASM, LGDP, LTHE; LASM, LGDP, LPHE; LIMR, LGDP, LTHE; and LIMR, LGDP and LPHE can be built. Different versions of each model can be obtained by adding dummies to represent structural breaks. In this section two models involving LASM, both of which include dummy variables representing the statistically significant structural breaks identified by the LM-2 test, are examined.[33] The model consisting of LASM, LGDP and LTHE includes an extra dummy for the year 1960 since as previously mentioned, additional categories of expenditure were included in the definition of total health expenditures.[34] A break was also included in 1958 for the model involving LASM, LGDP and LPHE to reflect the fact that beginning in July 1958, public health expenditures also included federal contributions to hospital insurance.[35]

Table 9 displays the results of pairwise Granger causality tests involving the variables of the two models. As for the 1960-1997 period, the null hypothesis that LGDP does not influence LASM at the 5% level of significance can be rejected. However, the null hypothesis that LASM does not influence LGDP cannot be rejected. Thus income has a significant effect on health status, but health status does not have a statistically significant impact on income. However, in contrast to the results for 1960-1997, in neither model does health spending appear to have a significant impact on health status at even the 10% level of significance. As far as the influence of health status on health spending is concerned, the results for Model 12A are fairly consistent with those for Model 9B, but those for Model 12B are not consistent with those for Model 9C. For the 1950-1997 period, the results suggest that there is no relationship of any kind between public health spending and health status. Public health spending does seem to depend on per capita income, though, since the null hypothesis that per capita income does not cause per capita public spending on health can be rejected at the 1% level of significance. But total health spending in Model 12A does not depend on income. Instead, the results suggest that both total and public health spending had a significant influence on per capita income during this period, an unexpected result which would appear to be primarily due to the extension of the sample period.[36] To the extent that health expenditures are correlated with total government expenditures

[33] Results for other models are available from the authors upon request. The analysis of impulse response functions of these models produced similar results.
[34] In contrast to the other dummy variables in the model, this dummy was defined to equal 1 in 1960 and zero in all other years, since it appears to cause only a level shift in LTHE.
[35] This dummy variable is defined to equal zero for $t \leq 1958$, and one for $t > 1958$.
[36] Of course, the dummy variables used in these models differ from those used in Models 9A-9C also.

during the 1950-1997 period, it is possible that this result reflects the positive influence of government spending on per capita GDP.

Thus once again, from a statistical point of view, one of the strongest observed relationships between the variables appears to be the influence of income on health status. For this sample period, health spending also seems to have an influence on per capita income. However, in contrast to the 1960-1997 period, there is no evidence that health spending has any direct impact on health status. To the extent that health spending influences income per capita, though, it will have an indirect impact on health status through that variable.

The generalized impulse responses for the two models are shown in Figure 12. These impulse response functions once again show that the system responds very slowly to shocks; indeed, the model consisting of LASM, LGDP and LTHE (panel A) does not seem to converge at all. In terms of the direction of the effects, most results are similar to those obtained previously. For example, although changes in per capita health expenditures, both total and public, still seem to increase age-standardized mortality in the short run, after several years age-standardized mortality begins to decrease. As well, in both models, positive shocks to LGDP seem to decrease LASM in the medium term, after a slight increase in the first years, while innovations in LGDP have a positive impact on LTHE and LPHE. However, with the exception of the impact of a GDP shock on per capita public health spending, most of the responses are very small in magnitude, and none of the responses exceed their 95% asymptotic confidence bounds. Thus even with the longer sample of data the estimated relationships between the variables do not appear to be very strong.

5.3 1926-1999

5.3.1 *Unit Root Tests*

For the 1926-1999 period, only two variables are examined as long time series were simply not available for the others. The results of the ADF-GLS and MZ_α tests for the infant mortality rate and real GNP per capita are presented in Table 10. The levels and first differences of the two variables are graphed in Figures 13 and 14 respectively.

As in the shorter sample periods, the ADF-GLS and MZ_α tests indicate that the infant mortality rate is integrated of order two. In the case of real GNP per capita, the two tests indicate that the null hypothesis of a unit root in the level of LGNP can be rejected at the 5% level of significance, which suggests that real GNP per capita is trend stationary with no unit roots. This result is somewhat surprising, since real GDP per capita was found to have at least one unit root in the shorter sample periods, and the tests do not allow for structural breaks.

Table 11 presents the results of the Perron and Rodriguez (2003) ADF-GLS and MZ_α tests allowing for one structural break. Given the results of the tests with no structural break, it is not surprising that LGNP is once again found to be trend stationary, albeit with a break this time. However, the two tests yield very different estimates of the break date – 1981 for the ADF-GLS test and 1936 for the MZ_α test. Compared to the shorter time periods, the order of integration indicated by these two tests is the same as for the 1960-1997 period but not the

1950-1997 period. In the latter period, real per capita GDP was I(1). The break dates are all different from those identified earlier.

For the infant mortality rate, both tests indicate that it is I(1) around a structural break. The break dates are again different: 1966 for the ADF-GLS test versus 1970 for the MZ_α test. The order of integration indicated by these tests is similar to that for the 1960-1997 period, but once again different from that for the 1950-1997 period. Again, the break dates are not the same.

Finally, the results of the LM tests in Table 12 indicate that both variables are I(0). The LM-2 test, which is to be preferred for this longer sample period because it allows for more breaks, indicates that breaks in LGNP occurred in 1948 and 1989. The first break date seems reasonable since it is soon after the end of World War II. However, it is less clear why 1989 is also a break date. For the infant mortality rate, the breaks occur in 1959 and 1977.

5.3.2 *VAR Analysis*

The finding of the LM-2 test that both LGNP and LIMR implies that it is appropriate to model the dynamic behaviour of the variables using a VAR model. Because there are so few variables involved, in this case the results for only one model are presented. The Granger causality tests for this VAR model are presented in Table 13. Not surprisingly, the results of the tests indicate that per capita income (this time measured by LGNP) does cause the infant mortality rate. In addition, the infant mortality rate causes per capita income. Note that this latter result conflicts with those for the shorter time periods, where the health status measure (LASM) did not cause income per capita. While it is tempting to interpret this result as evidence that improvements in health have contributed to economic growth, it would be premature to do so without further research. Indeed, in a simple bivariate model, one cannot be sure whether the observed causality is the result of a direct relationship between the two variables, or an indirect relationship mediated through other variables that are not included in the model. In addition, until the impulse responses are examined, nothing can be said about the sign of the relationship between the variables.

The generalized impulse responses are presented in Figure 15. In this two-variable VAR, adjustments to shocks are very slow – after 100 periods, the two variables still have not converged. Also, the initial impact of the shock to the LGNP equation on the infant mortality rate is in the wrong direction – it increases the infant mortality rate, though only by about 1%, and after five years has reduced the infant mortality rate to about 1% below its initial level. However, the response to a shock to LIMR seems to be in the right direction – an increase in the infant mortality rate causes LGNP to fall. Interestingly, the short-run impact on LGNP of a shock to the LIMR equation is much larger than the effects of shocks to the LASM equation on LGDP for the shorter sample periods: five years after the increase in the infant mortality rate, LGNP has decreased by almost 4%, as compared to changes of less than 1% for the previous models. But once again the impulse responses lie within the 95% confidence bounds, indicating that they are not measured with precision. Thus although the causality tests indicate that there is a statistically significant relationship between the infant mortality rate and per capita income over the 1926-1999 time period, the impulse response analysis suggests that the growth in income over this period did not have a very strong effect on the infant mortality rate.

6. Conclusion

Several conclusions can be drawn from this analysis. First, the results of the unit root tests are clearly sensitive to the choice of sample period and to the choice of test. For example, real per capita GDP was found to be integrated of both order one and order zero, depending on the sample period and on the unit root test used. However, the results of the LM test allowing for one or two structural breaks were generally consistent from one sample period to another, suggesting that most series were I(0) with one or two breaks in trend. The existence of these breaks could explain why the ADF-GLS test often indicated that variables were I(2).[37]

Second, the analysis of VAR models through causality testing and an examination of the impulse response functions of a number of VAR models constructed using the variables indicated that there are some statistically significant relationships between the variables. The most consistent finding was that income per capita does influence health status. For the 1960-1997 sample period, there was some evidence that per capita public health spending also influences health status, but these results were not duplicated for the 1950-1997 sample period. Furthermore, the impulse response analyses suggest that historically, the effects on health status of changes in health spending were small in magnitude. In most models, it took more than fifty years for the effect of a shock to dissipate, while other models did not even converge. As well, most impulse responses were not statistically significant. Overall, the results of these dynamic analyses suggest that a specific statistical relationship between these three variables alone – health status, per capita GDP, and per capita health spending – may not exist in Canada.

One possible explanation for this apparent lack of a relationship may be threshold effects. For example, at lower levels of health spending and health status, the association between health expenditures and health outcomes may be strong, but at high levels of population health status, the association may be weaker. As Canada's population health status is considered relatively high, it would experience lower returns to health spending. This reasoning also applies to the relationship between GDP and health outcomes. In non-developed countries where health status is lower, one would expect that an increase in GDP would have a greater impact on health status than in developed countries.

Alternatively, the observed lack of convergence and inconsistent results in the VAR models may be an indication of model misspecification. To verify this, it would be worthwhile developing a more complete structural model of the interactions between per capita health spending, per capita GDP, and health outcomes; that is, a model that includes more variables. However, the sample period available for the estimation of such a model would necessarily be more restricted than that employed here. It would also be worthwhile to further investigate the nature of the structural breaks identified by the unit root tests, and their impact on health status and health spending in Canada. To the extent that these breaks are related to policy changes, they provide another channel through which policy may have influenced health status, health spending, and income per capita in Canada. Certainly, the relationship between health outcomes, health spending and per capita GDP is not straightforward. Therefore, further research is needed to determine the exact nature of this relationship.

[37] The results in appendix B show that the same is true for other OECD countries.

REFERENCES

Auster, R., I. Leveson, and D. Sarachek (1969) "The Production of Health, an Exploratory Study." *Journal of Human Resources* 4(4), 411-436.

Baldacci, E., M.T. Guin-Siu, and L. de Mello (2002) "More on the Effectiveness of Public Spending on Health Care and Education: A Covariance Structure Model." IMF Working Paper WP/02/90, Washington, International Monetary Fund.

Berger, M.C. and J.P. Leigh (1989) "Schooling, Self-Selection, and Health." *Journal of Human Resources* 24(3), 433-455.

Berger, M. C., and J. Messer (2002) "Public Financing of Health Expenditures, Insurance, and Health Outcomes." *Applied Economics* 34(17), 2105-2113.

Blomqvist, Å.G., and R.A.L. Carter (1997) "Is Health Care Really a Luxury?" *Journal of Health Economics* 16(2), 207-229.

CIHI (2001) *National Health Expenditure Trends, 1975-2001*. Canadian Institute for Health Information, Ottawa.

CIHI (2004) *National Health Expenditure Trends, 1975-2004*. Canadian Institute for Health Information, Ottawa.

Crémieux, P.-Y., P. Ouellette, and C. Pilon (1999) "Health Care Spending as Determinants of Health Outcomes." *Health Economics* 8(7), 627-639.

Crémieux, P.-Y., M.-C. Meilleur, P. Ouellette, P. Petit, M. Zelder, and K. Potvin (2005a) "Public and Private Pharmaceutical Spending as Determinants of Health Outcomes in Canada." *Health Economics* 14(2), 107-116.

Crémieux, P.-Y., M.-C. Meilleur, P. Ouellette, P. Petit, M. Zelder, and K. Potvin (2005b) "Erratum: 'Public and Private Pharmaceutical Spending as Determinants of Health Outcomes in Canada.'" *Health Economics* 14(2), 117.

Deussing, M.-A. (2003) "An Empirical Analysis of the Relationship between Public Health Spending and Self-Assessed Health Status: An Ordered Probit Model." M.A. Major Paper, Department of Economics, University of Ottawa, Ottawa.

Di Matteo, L., and R. Di Matteo (1998) "Evidence on the Determinants of Canadian Provincial Government Health Expenditures: 1965-1991." *Journal of Health Economics* 17(2), 211-228.

Elliott, G., T.J. Rothenberg, and J.H. Stock (1996) "Efficient Tests for an Autoregressive Unit Root." *Econometrica* 64(4), 813-836.

Filmer, D., and L. Pritchett (1999) "The Impact of Public Spending on Health: Does Money Matter?" *Social Science and Medicine* 49(10), 1309-1323.

Gerdtham, U.-G., and B. Jönsson (2000) "International Comparisons of Health Expenditure: Theory, Data and Econometric Analysis." In A. J. Culyer and J. P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1A. Amsterdam: Elsevier, 11-53.

Gerdtham, U.-G., and M. Löthgren (2000) "On Stationarity and Cointegration of International Health Expenditure and GDP." *Journal of Health Economics* 19(4), 461-475.

Gerdtham, U.-G., and M. Löthgren (2002) "New Panel Results on Cointegration of International Health Expenditure and GDP." *Applied Economics* 34(3), 1679-1686.

Gerdtham, U.-G., J. Søgaard, F. Andersson, and B. Jönsson (1992) "An Econometric Analysis of Health Care Expenditure: A Cross-Section Study of the OECD Countries." *Journal of Health Economics* 11(1), 63-84.

Gravelle, H. S. E., and M. E. Backhouse (1987) "International Cross-Section Analysis of the Determination of Mortality." *Social Science and Medicine* 25(5), 427-441.

Gregory, A. W. and B. E. Hansen (1996) "Residual-Based Tests for Cointegration in Models with Regime Shifts." *Journal of Econometrics* 70(1), 99-126.

Grossman, M. (1972) *The Demand for Health: A Theoretical and Empirical Investigation*. New York: Columbia University Press for the National Bureau of Economic Research.

Hanratty, M. J. (1996) "Canadian National Health Insurance and Infant Health." *American Economic Review* 86(1), 276-284.

Hansen, P., and A. King (1996) "The Determinants of Health Care Expenditure: A Cointegration Approach." *Journal of Health Economics* 15(1), 127-137.

Hansen, P., and A. King (1998) "Health Care Expenditure and GDP: Panel Data Unit Root Test Results – Comment." *Journal of Health Economics* 17(3), 377-381.

Harvey, D. I., S. J. Leybourne and P. Newbold (2001) "Innovational Outlier Unit Root Tests with an Endogenously Determined Break in Level", Oxford Bulletin of Economics and Statistics 63(5), 559-575.

Hitiris, T., and J. Posnett (1992) "The Determinants and Effects of Health Expenditure in Developed Countries." *Journal of Health Economics* 11(2), 173-181.

Im, K.S., J. Lee, and M. Tieslau (2005) "Panel LM Unit-Root Tests with Level Shifts." *Oxford Bulletin of Economics and Statistics* 67(3), 393-419.

Im, K.S., M.H. Pesaran, and Y. Shin (2003) "Testing for Unit Roots in Heterogeneous Panels." *Journal of Econometrics* 115(1), 53-74.

Jackson, H., and A. McDermott (2004) "Health-Care Spending: Prospect and Retrospect." Finance Canada Analytical Note, Finance Canada, Ottawa, January.

Jee, M., and Z. Or (1998) "Health Outcomes in OECD Countries: A Framework of Health Indicators for Outcome-Oriented Policymaking." *Labour Market and Social Policy- Occasional Papers No. 36*, OECD, Paris.

Jewell, T., J. Lee, M. Tieslau, and M.C. Strazicich (2003) "Stationarity of Health Expenditures and GDP: Evidence from Panel Unit Root Tests with Heterogeneous Structural Breaks". *Journal of Health Economics* 22(2), 313-323.

Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Juselius, K. (2005) "Chapter 6: Deterministic Components in the I(1) Model." In *The Cointegrated VAR Model: Econometric Methodology and Macroeconomic Applications* (preliminary title). Revised February 27, 2005. Institute of Economics, University of Copenhagen. Available at http://www.econ.ku.dk/okokj/ .

Kee, G.-S. (2001) "An Empirical Analysis of Canadian Public Health Care Spending and Health: 1975 to 1996." Master's Thesis, Calgary, University of Calgary.

Knowles, S., and P.D. Owen (1995) "Health Capital and Cross-Country Variation in Income per Capita in the Mankiw-Romer-Weil Model", *Economics Letters* 48(1), 99-106.

Koop, G., M.H. Pesaran, and S.M. Potter (1996) "Impulse Response Analysis in Nonlinear Multivariate Models." *Journal of Econometrics* 74(1), 119-47.

Leacy, F.H., ed. (1983) *Historical Statistics of Canada*. Ottawa: Statistics Canada, Catalogue 11-516-XIE. Available on the internet at http://www.statcan.ca:80/english/freepub/11-516-XIE/free.htm.

Lee, J., and M. Strazicich (2001) "Break Point Estimation and Spurious Rejections with Endogenous Unit Root Tests." *Oxford Bulletin of Economics and Statistics* 63(5), 535-558.

Lee, J., and M. Strazicich (2003) "Minimum Lagrange Multiplier Unit Root Test with Two Structural Breaks." *Review of Economics and Statistics* 85(4), 1082-1089.

Lee, J., and M. Strazicich (2004) "Minimum LM Unit Root Test with One Structural Break" Working Paper, Department of Economics, Appalachian State University.

Lichtenberg, F. R. (2004) "Sources of U.S. Longevity Increase, 1960-2001." *Quarterly Review of Economics and Finance* 44(3), 369-389.

Lise, Jeremy (2000) “Changes in Life Expectancy and Mortality Rates: 1926-1966.” Ottawa: Health Canada, Applied Research and Analysis Directorate, Information, Analysis and Connectivity Branch, March 28.

Maddala, G.S., and I.-M. Kim (1998) *Unit Roots, Cointegration, and Structural Change.* Cambridge: Cambridge University Press.

McCoskey, S.K., and T.M. Selden (1998) ‘Health Care Expenditures and GDP: Panel Data Unit Root Test Results.” *Journal of Health Economics* 17(3), 369-376.

McDonald, S. and J. Roberts (2002) “Growth and Multiple Forms of Human Capital in an Augmented Solow Model: A Panel Data Investigation”, *Economics Letters* 74(2), 271-76.

Morrison, D. F. (1967) *Multivariate Statistical Methods*. New York: McGraw-Hill Book Company.

Ng, S. and P. Perron (2001) “Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power.” *Econometrica* 69(6), 1519-1554.

Okunade, A.A., and M.C. Karakus (2001) “Unit Root and Cointegration Tests: Time-Series versus Panel Estimates for International Health Expenditure Models.” *Applied Economics* 33(9), 1131-1137.

Or, Z. (2000a) “Exploring the Effects of Health Care on Mortality across OECD Countries.” *Labour Market and Social Policy Occasional- Papers No. 46*, Paris, Organisation for Economic Cooperation and Development.

Or, Z. (2000b) “Determinants of Health Outcomes in Industrialised Countries: A Pooled, Cross-Country, Time-Series Analysis.” *OECD Economic Studies: No. 30*, 2000/1, Paris, Organisation for Economic Cooperation and Development.

Perron, P., and S. Ng (1996) “Useful Modifications to Some Unit Root Tests with Dependent Errors and Their Local Asymptotic Properties.” *Review of Economic Studies* 63(3), 435-463.

Perron, P., and G. Rodriguez (2003) “GLS Detrending, Efficient Unit Root Tests and Structural Change.” *Journal of Econometrics* 115(1), 1-27.

Pesaran, M.H. and Y. Shin (1998) “Generalized Impulse Response Analysis in Linear Multivariate Models.” *Economics Letters* 58(1), 17-29.

Preston, S.H. (1975) “The Changing Relation between Mortality and Level of Economic Development.” *Population Studies* 29(2), 231-248.

Rivera, B. (2001) “The Effects of Public Health Spending on Self-Assessed Health Status: An Ordered Probit Model.” *Applied Economics* 33(10), 1313-1319.

Roberts, J. (1999) “Sensitivity of Elasticity Estimates for OECD Health Care Spending: Analysis of a Dynamic Heterogeneous Data Field.” *Health Economics* 8(5), 459-472.

Thornton, J. (2002) “Estimating a Health Production Function for the US: Some New Evidence.” *Applied Economics* 34(1), 59-62.

Vogelsang, T. J. (1999) “Two Simple Procedures for Testing for a Unit Root When There Are Additive Outliers.” *Journal of Time Series Analysis* 20(2), 237-252.

Figure 1
Infant Mortality Rate, Age-Standardized Mortality Rate, Perinatal Mortality Rate, Life Expectancy at Birth and Life Expectancy at 65 , Canada, 1950 to 1997 (1950=100)

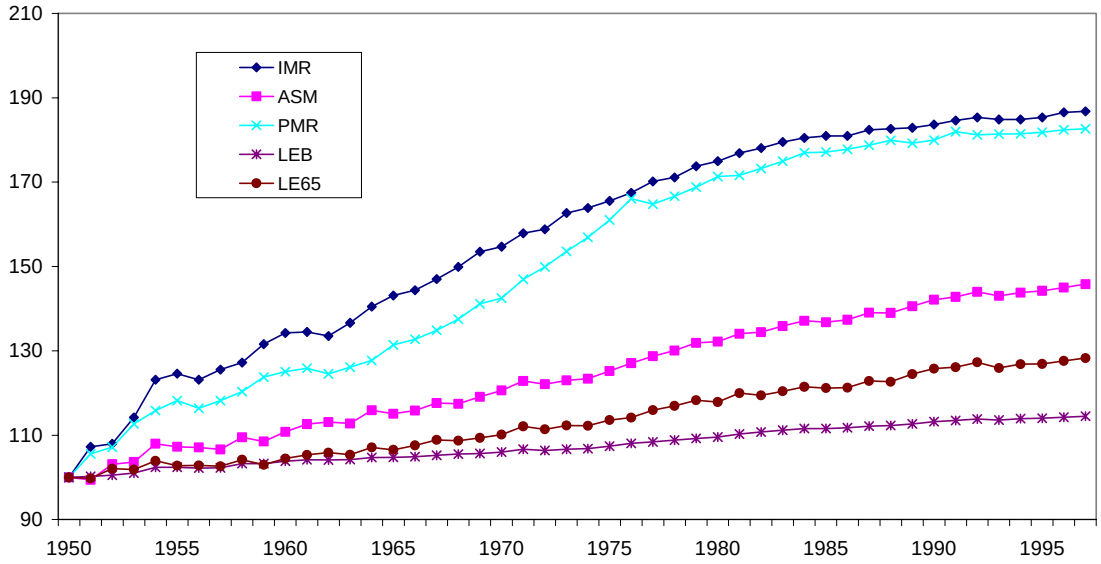


Figure 2
Infant Mortality Rate, Perinatal Mortality Rate, Life Expectancy at Birth and Life Expectancy at 65, Canada, 1926 to 1997 (1926=100)

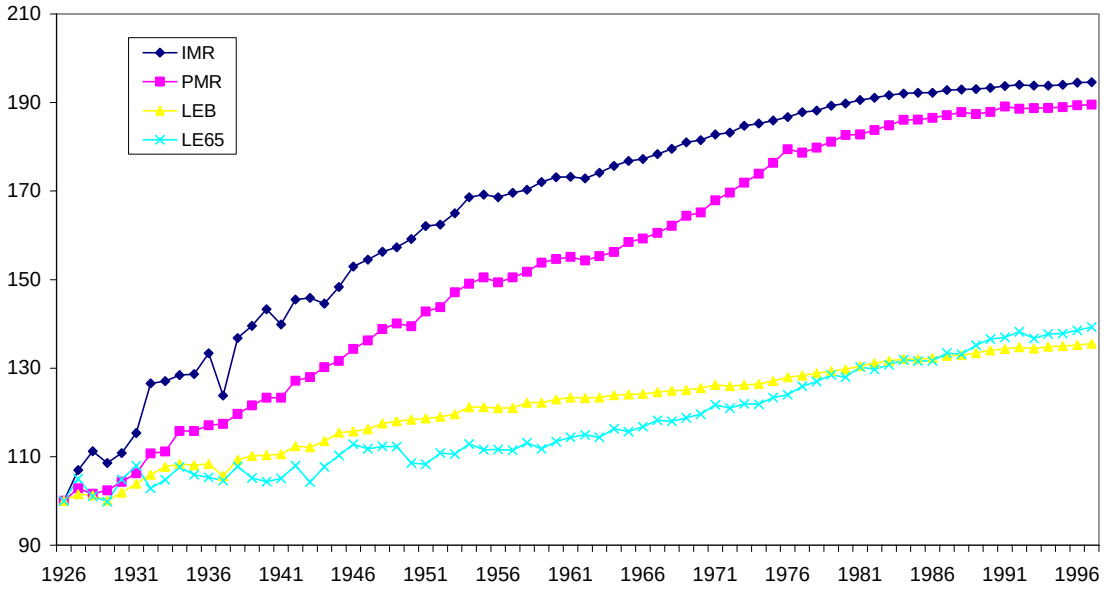


Figure 3
Summary Indicator of Health Status for Canada, 1950 to 1997

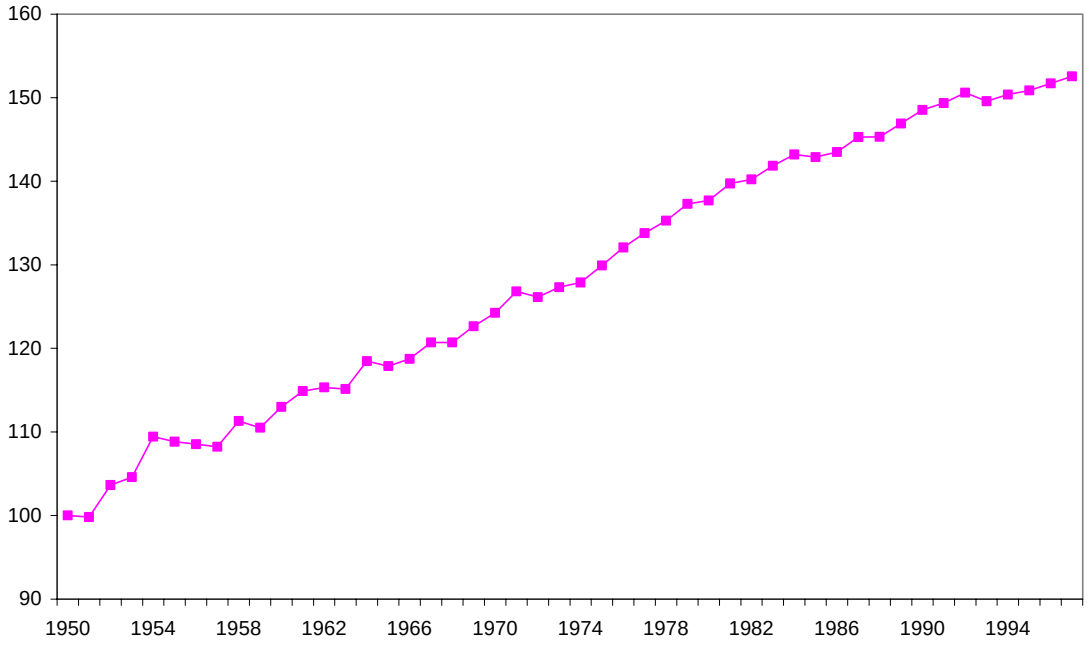


Figure 4
Public/Private Health Expenditures, Real per Capita, Age-Adjusted, Canada, 1976 to 2001

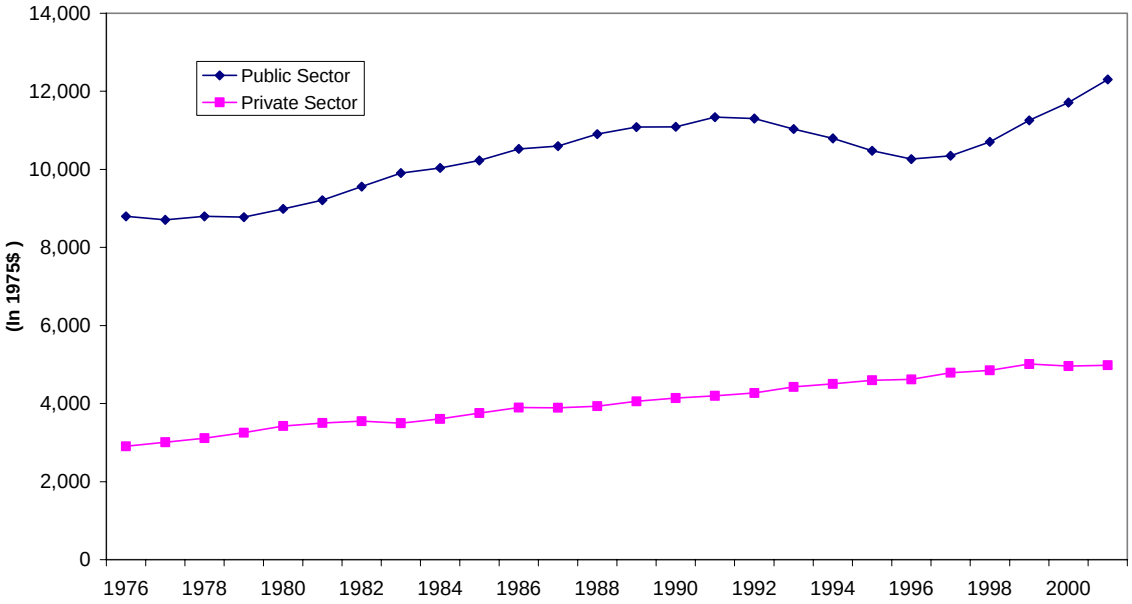


Figure 5
Total Real per Capita Health Expenditures, Canada, 1945 to 1999

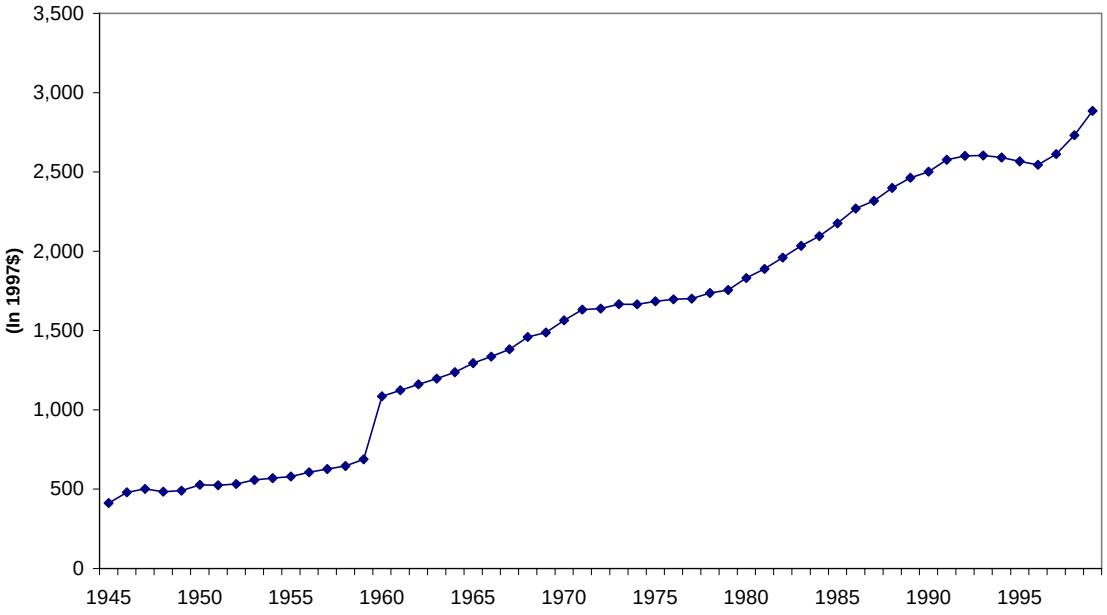


Figure 6
Real GNP per Capita and Real GDP per Capita, Canada, 1926 to 1999

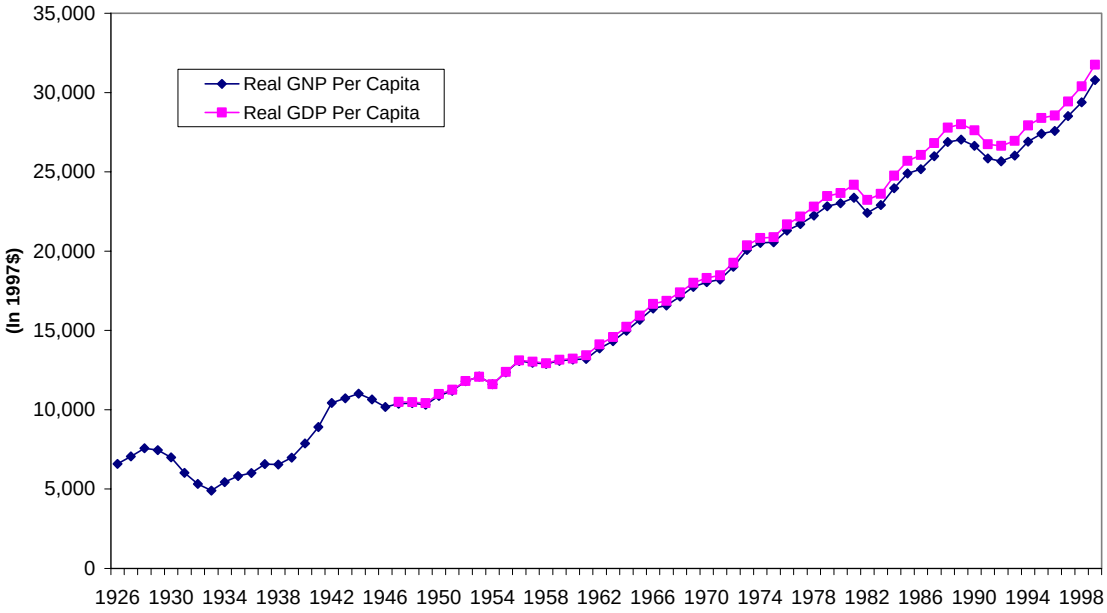


Figure 7a
LHS, 1960 to 1997

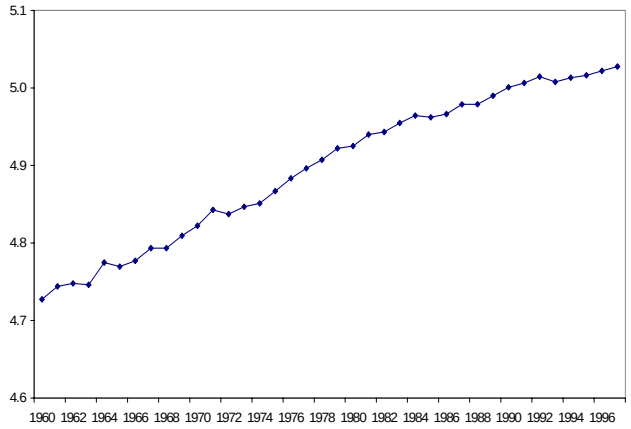


Figure 7b
LTHE, 1960 to 1997

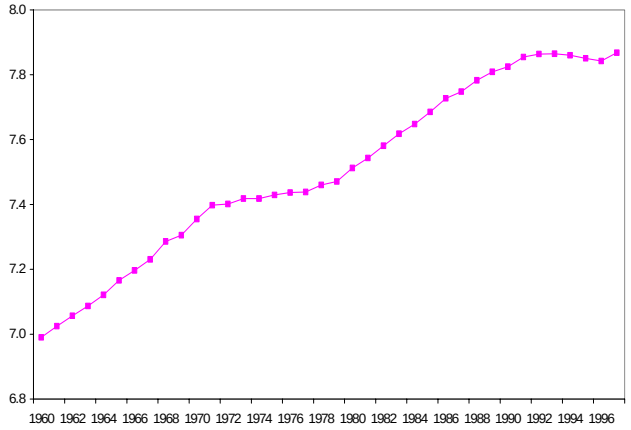


Figure 7c
LGDP, 1960 to 1997

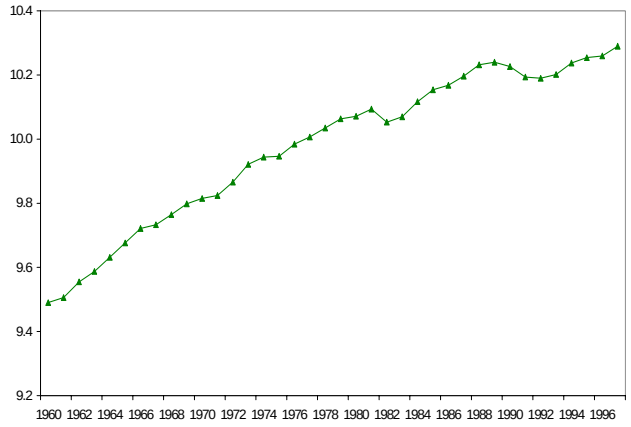


Figure 7d
LASM, 1960 to 1997

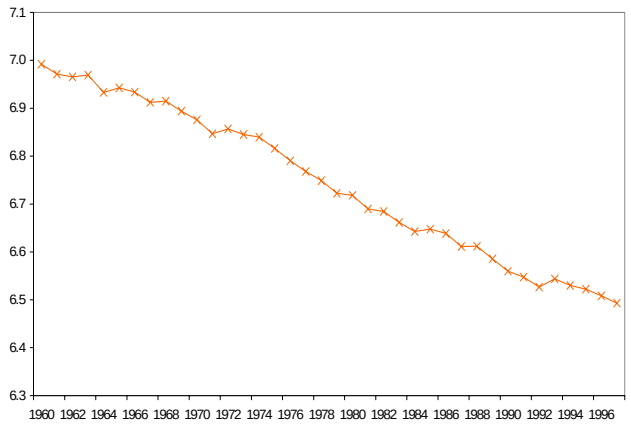


Figure 7e
LIMR , 1960 to 1997

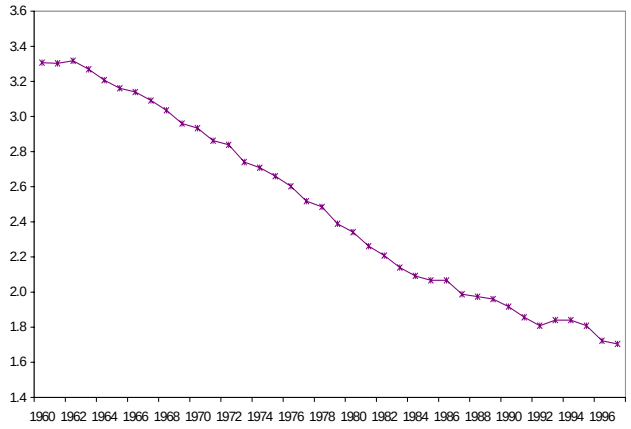


Figure 7f
LPHE, 1960 to 1997

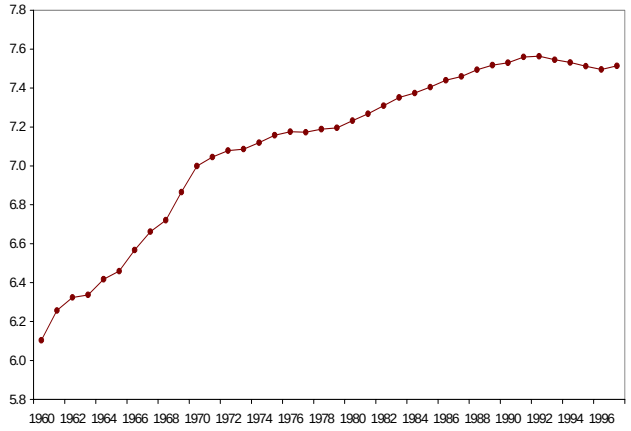


Figure 8a
First Difference of LHS, 1961 to 1997

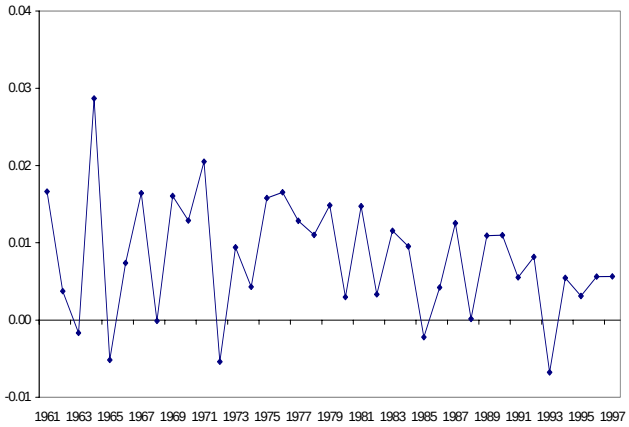


Figure 8b
First Difference of LTHE, 1961 to 1997

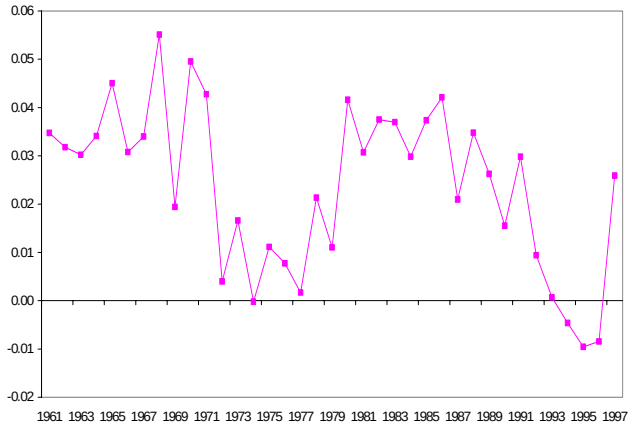


Figure 8c
First Difference of LGDP, 1961 to 1997

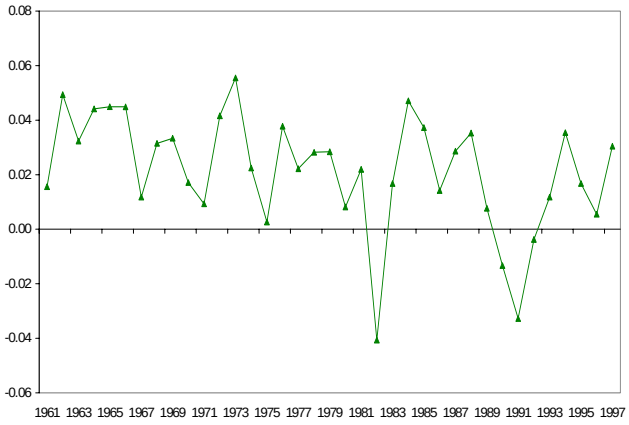


Figure 8d
First Difference of LASM, 1961 to 1997

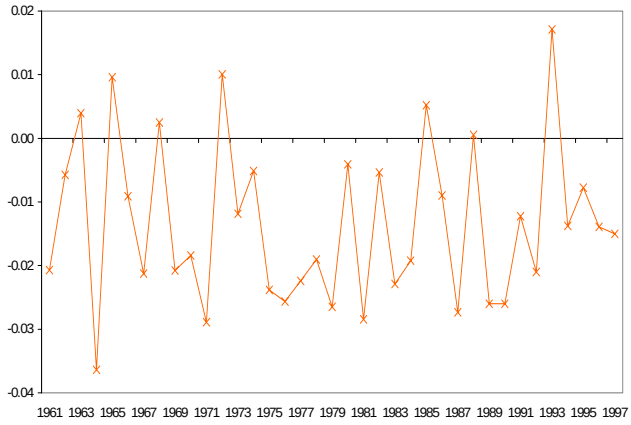


Figure 8e
First Difference of LIMR , 1961 to 1997

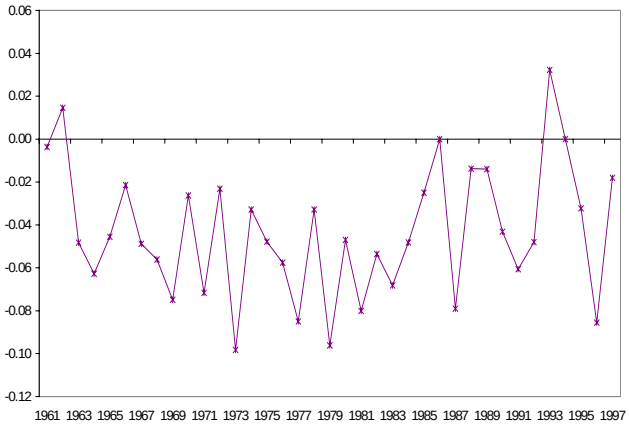


Figure 8f
First Difference of LPHE, 1961 to 1997

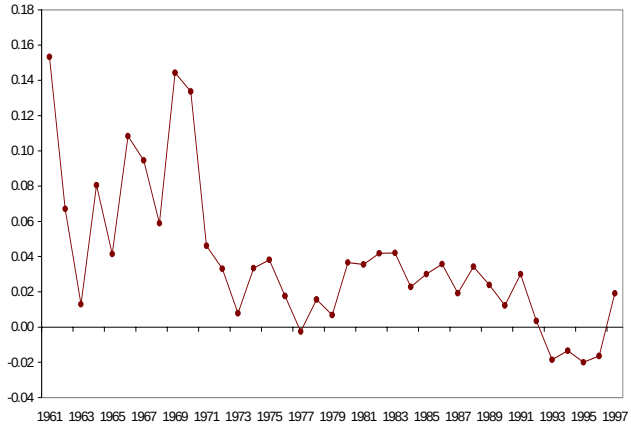
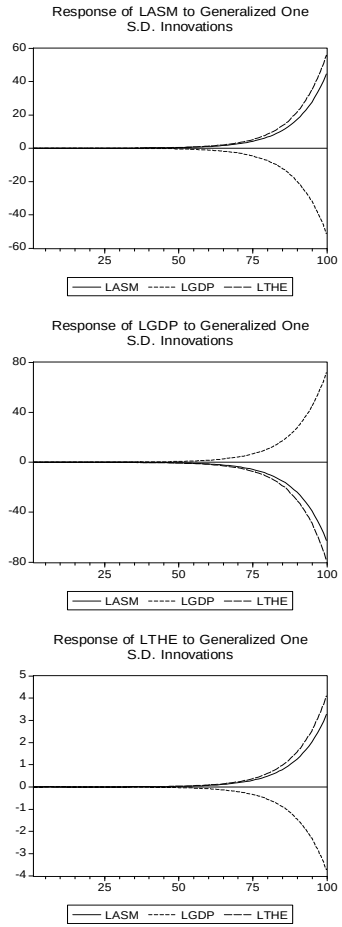
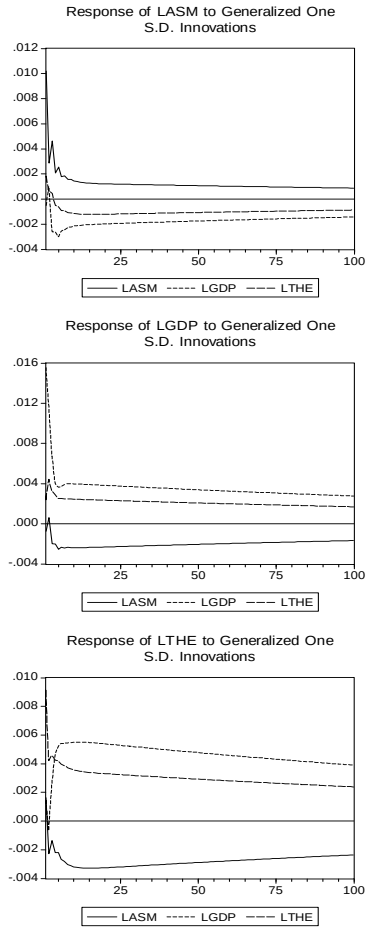


Figure 9
Generalized Impulse Response Functions for Selected VAR Models, 1960-1997

A. LASM, LGDP, LTHE, and dummies for 1979, 1980, and 1982 following results of LM-1 test (2 lags)



B. LASM, LGDP, LTHE, and dummies for 1971, 1975, 1980, 1984, 1989, and 1993 following results of LM-2 test (2 lags)



C. LASM, LGDP, LPHE, and dummies for 1976, 1979, and 1980 following results of LM-1 test (2 lags)

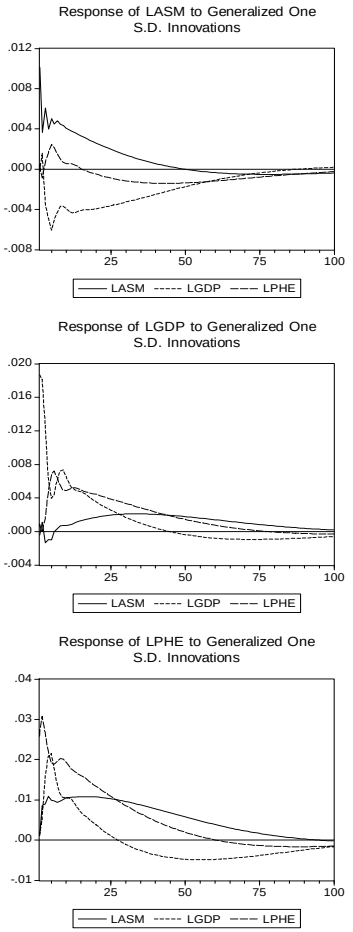


Figure 10a
LHS, 1950 to 1997

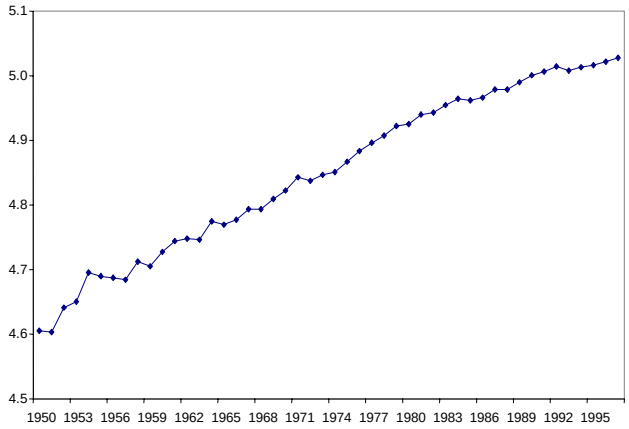


Figure 10b
LTHE, 1950 to 1997

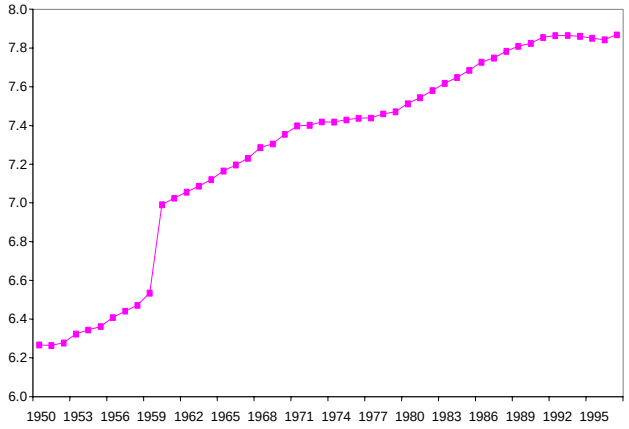


Figure 10c
LGDP, 1950 to 1997

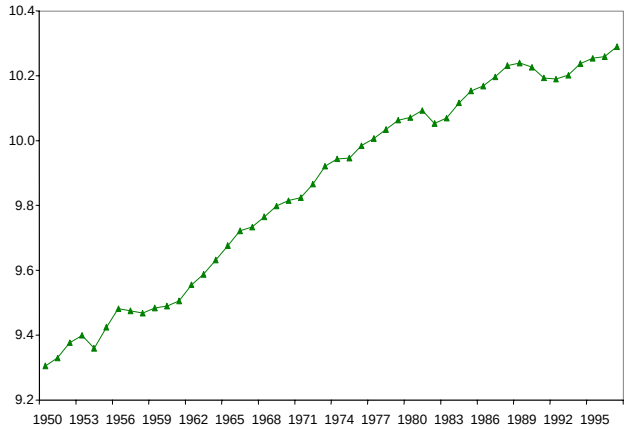


Figure 10d
LASM, 1950 to 1997

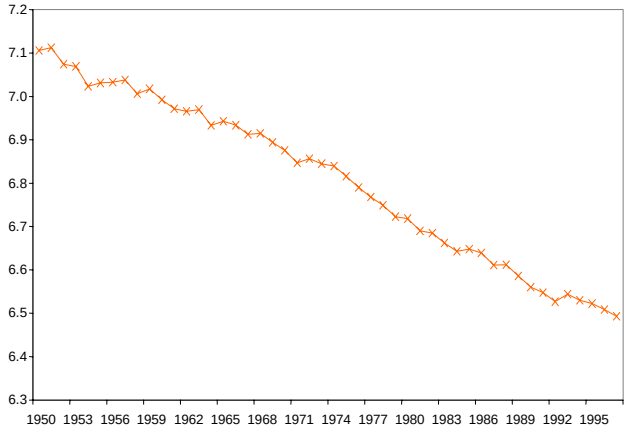


Figure 10e
LIMR, 1950 to 1997

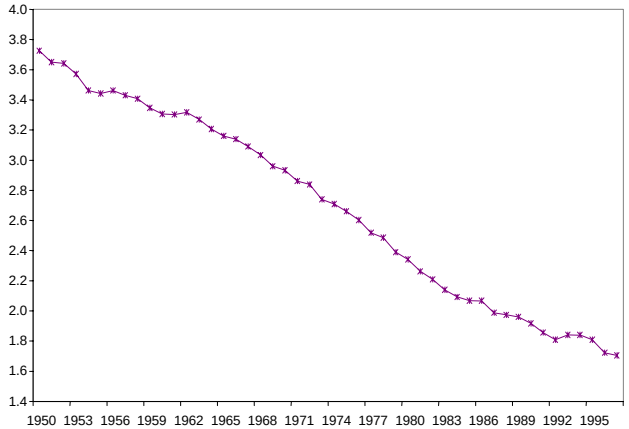


Figure 10f
LPHE, 1950 to 1997

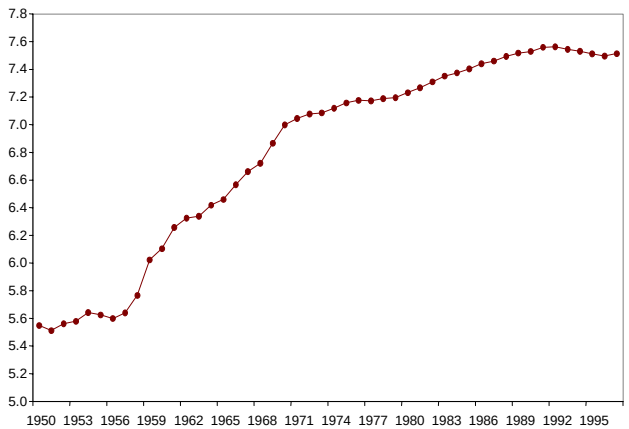


Figure 11a
First Difference of LHS, 1951 to 1997

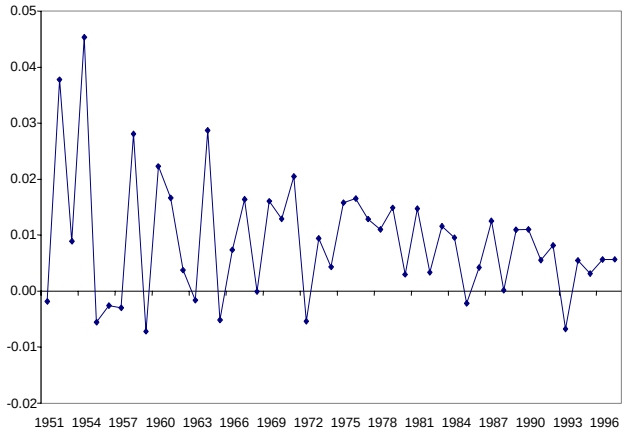


Figure 11b
First Difference of LTHE, 1951 to 1997

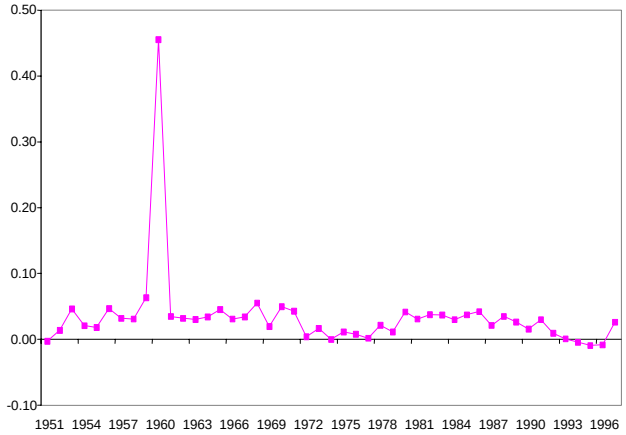


Figure 11c
First Difference of LGDP, 1951 to 1997

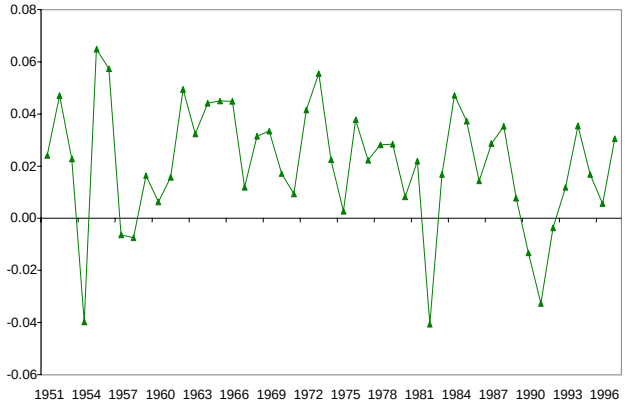


Figure 11d
First Difference of LASM, 1951 to 1997

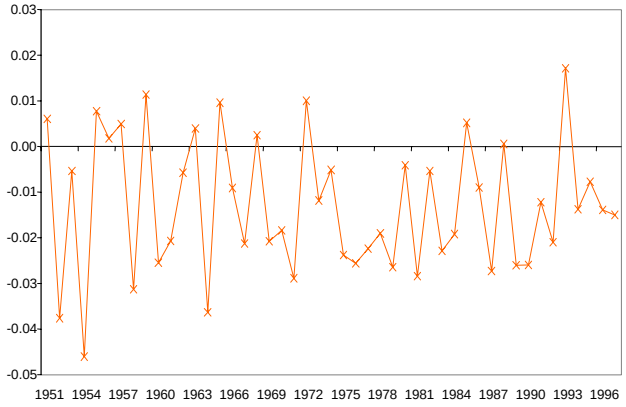


Figure 11e
First Difference of LIMR, 1951 to 1997

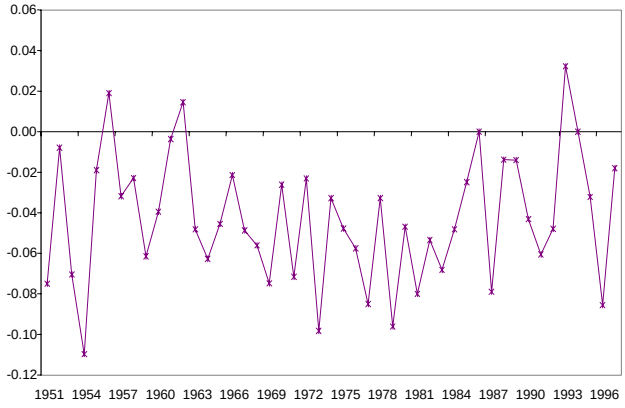


Figure 11f
First Difference of LPHE, 1951 to 1997

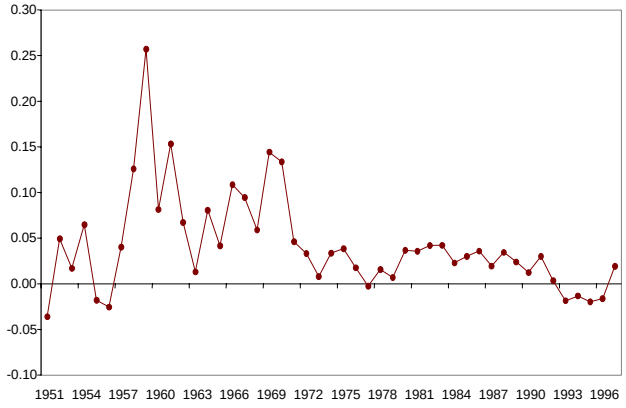
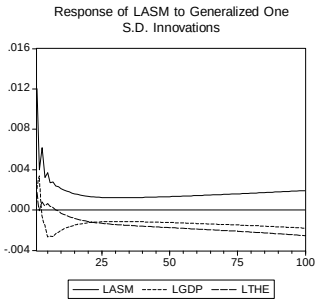
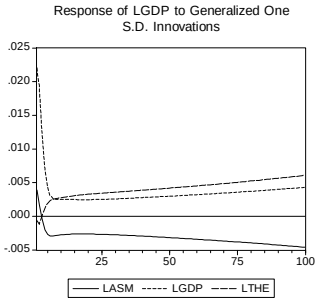
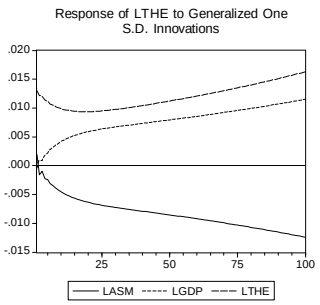


Figure 12
Generalized Impulse Response Functions for Selected VAR Models, 1950-1997

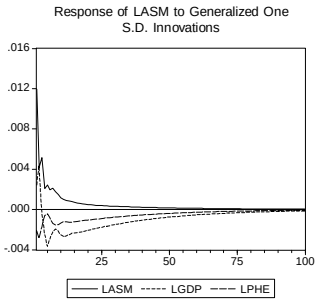
A. LASM, LGDP, LTHE and dummies for 1970, 1974, 1987, and 1992 following results of LM-2 test; pulse dummy for 1960 (2 lags)

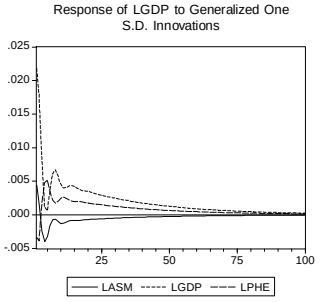






B. LASM, LGDP, LPHE and dummies for 1967, 1970, 1974, 1978, and 1987 following results of LM-2 test; additional dummy for 1958 (2 lags)





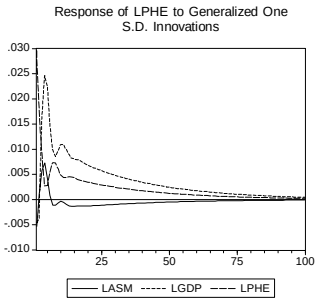


Figure 13a
LGNP, 1926 to 1999

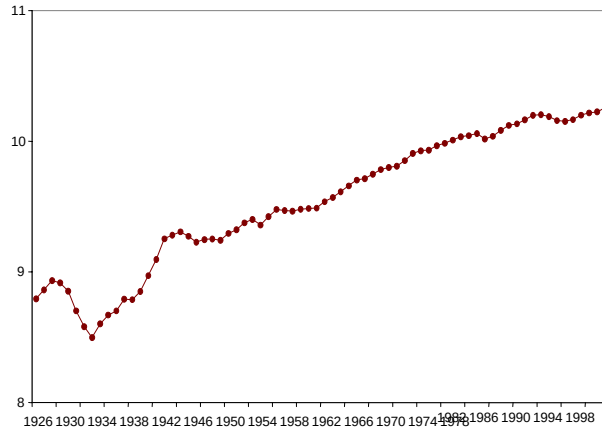


Figure 13b
LIMR, 1926 to 1999

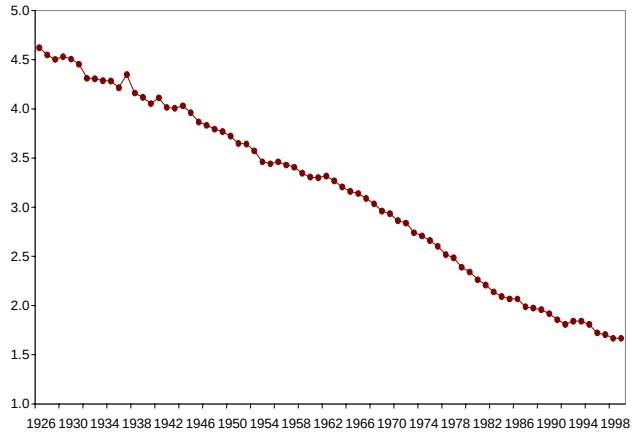


Figure 14a
First Difference of LGNP, 1927 to 1999

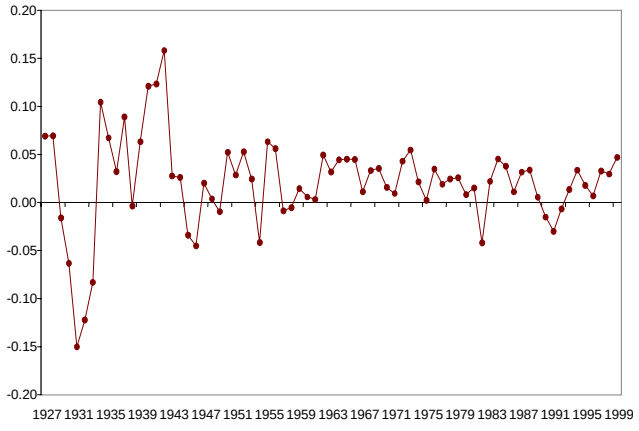


Figure 14b
First Difference of LIMR, 1927 to 1999

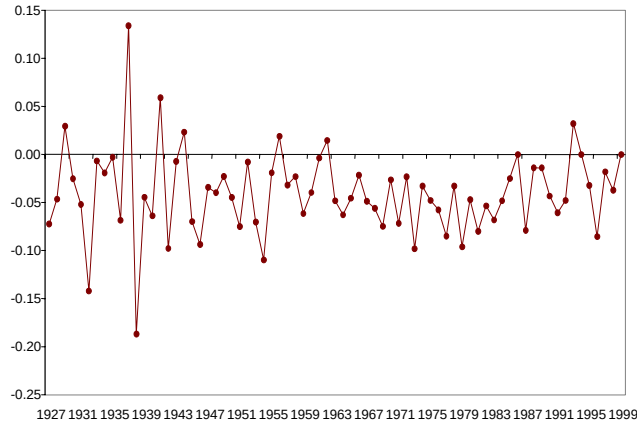


Figure 15
Generalized Impulse Response Functions for a VAR Model, 1926-1999

LIMR, LGNP, and dummies for 1948, 1959, 1977,
and 1989 following results of LM-2 test (5 lags)

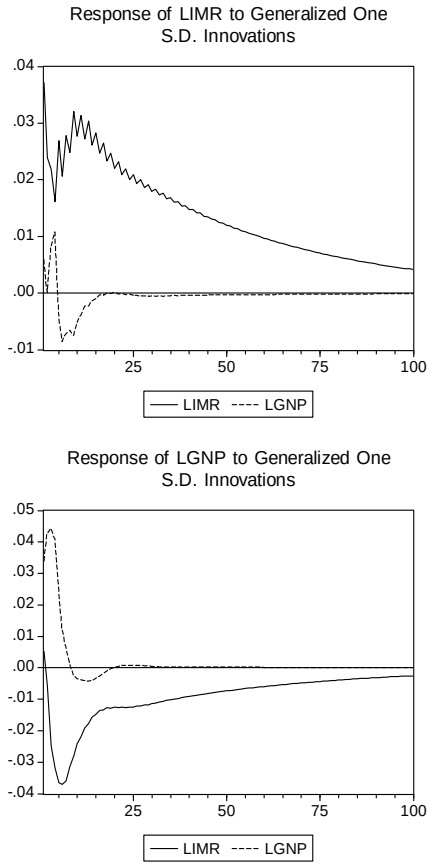


Table 1. Unit root test results for Canada: Results of previous studies

Study	Unit root test	Sample period	Order of integration	
			HE	GDP
Hansen and King (1996)	ADF	1960-1987	I(0)	I(1)
Blomqvist and Carter (1997)	PP	1960-1991	I(1)	I(1)
McCoskey and Selden (1998)	IPS	1960-1987	I(0)	I(1)
Gerdtham and Löthgren (2000)	ADF, IPS, KPSS	1960-1997	I(1)	I(1)
Okunade and Karakus (2001)	ADF, PP, IPS	1960-1997	I(2)	I(2), I(1)
Gerdtham and Löthgren (2002)	ADF, IPS	1960-1997	I(1)	I(1)
Jewell et al. (2003)	Im et al. (2005)	1960-1997	I(1)	I(0)

Note: Only Hansen and King (1996), Blomqvist and Carter (1997), and Okunade and Karakus (2001) present the results of unit root tests on both the levels and differences of variables.

Table 2. Results of ADF-GLS and MZ_α unit root tests with no structural break, 1960-1997, Canada

Variable	k [d]	T [e]	ADF-GLS	MZ_α
LGDP	0	37	-1.060	-1.531
ΔLGDP	0	36	-4.154[c]	-17.060[c]
LTHE	1	36	-1.914	-11.1014
ΔLTHE	1	35	-1.789[a]	-6.327[a]
Δ^2 LTHE	2	33	-2.531[b]	-9.532[c]
LPHE	2	35	-1.016	-5.583
ΔLPHE	8	28	-0.003	-53.941[c]
Δ^2 LPHE	1	34	-7.146[c]	
LHS	0	37	-1.382	-4.442
ΔLHS	6	30	0.065	-0.475
Δ^2 LHS	0	35	-12.702[c]	-19.012[c]
LASM	1	36	-2.075	-6.938
ΔLASM	1	35	-4.414[c]	-14.726[c]
LIMR	0	37	-1.116	-2.872
ΔLIMR	4	32	-0.856	-0.790
Δ^2 LIMR	0	35	-10.551[c]	-15.703[c]

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test using MAIC. Maximum lag length was 9.
[e] Number of observations for ADF-GLS test. For the MZ_α test, $T = 37$.

Table 3. Results of ADF-GLS and MZ_α unit root tests with one structural break, 1960-1997, Canada

Variable	k [d]	T [e]	ADF-GLS	Tb[f]	MZ_α	Tb
LGDP	1	36	-4.032[a]	1975	-26.228[b]	1975
LTHE	1	36	-2.461	1978	-16.883	1967
ΔLTHE	1	35	-3.128	1981	-12.509	1981
LPHE	7	30	-3.505	1966	-104.648[c]	1966
ΔLPHE	1	35	-4.135[a]	1979		
LHS	1	36	-3.832	1980	-16.686	1980
ΔLHS	1	35	-6.439[a]	1975	-23.983[a]	1981
LASM	1	36	-3.583	1978	-16.501	1978
ΔLASM	1	35	-5.910[c]	1975	-21.600[a]	1975
LIMR	1	36	-2.644	1976	-11.953	1976
ΔLIMR	1	35	-6.016[c]	1983	-25.945[b]	1981

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test using MAIC. Maximum lag length was 9.
[e] Number of observations for ADF-GLS test.
[f] Time of break.

Table 4. Results of LM unit root tests with one or two structural breaks, 1960-1997, Canada

Variable	k[d]	LM-1 test	Tb[e]	k	LM-2 test	Tb
LGDP	7	-5.633[c]	1980	5	-6.847[c]	1980 1989
LTHE	3	-5.781[c]	1982	3	-7.051[c]	1971 1984
LPHE	8	-5.040[b]	1976	8	-12.508[c]	1976 1986
LHS	0	-4.436[a]	1977	0	-5.292	1976 1993
ΔLHS				0	-8.963[c]	1974 1976[f]
LASM	9	-5.139[b]	1979	0	-5.472[a]	1975 1993
LIMR	9	-4.778[b]	1977	6	-5.566[a]	1977 1992

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length. Maximum lag length was 9.
[e] Time of break.
[f] Break is not significant at the 10% level.

Table 5. Results of Causality Tests, 1960-1997, Canada

VAR Model 9A:			
LGDP → LASM	14.986 (0.0006)	LASM → LGDP	1.900 (0.3867)
LTHE → LASM	3.175 (0.2045)	LASM → LTHE	3.229 (0.1990)
LGDP → LTHE	3.708 (0.1566)	LTHE → LGDP	0.857 (0.6516)
VAR Model 9B:			
LGDP → LASM	6.687 (0.0353)	LASM → LGDP	0.719 (0.6980)
LTHE → LASM	0.070 (0.9658)	LASM → LTHE	6.459 (0.0396)
LGDP → LTHE	9.900 (0.0071)	LTHE → LGDP	4.239 (0.120)
VAR Model 9C:			
LGDP → LASM	13.257 (0.0013)	LASM → LGDP	2.056 (0.3578)
LPHE → LASM	4.683 (0.0962)	LASM → LPHE	8.294 (0.0158)
LGDP → LPHE	8.274 (0.0160)	LPHE → LGDP	3.774 (0.1516)

Note: Arrows indicate the direction of causality under the alternative hypothesis; for all tests, the null hypothesis is that no relationship exists. P-values are in parentheses.

Table 6. Results of ADF-GLS and MZ_α unit root tests with no structural break, 1950-1997, Canada

Variable	k [d]	T [e]	ADF-GLS	MZ_α
LGDP	1	46	-1.566	-6.245
ΔLGDP	9	37	-0.924	-0.446
Δ^2 LGDP	0	45	-8.643[c]	-22.865[c]
LTHE	0	47	-1.063	-2.556
ΔLTHE	7	39	-1.532	-10.474[b]
Δ^2 LTHE	0	45	-10.817[c]	
LPHE	5	42	-1.814	-28.182[c]
ΔLPHE	4	42	-1.175	-4.783
Δ^2 LPHE	3	42	-5.832[c]	-13.172[c]
LHS	1	46	-1.599	-3.890
ΔLHS	6	40	-0.980	-0.246
Δ^2 LHS	0	45	-16.422[c]	-30.636[c]
LASM	1	46	-1.808	-6.076
ΔLASM	7	39	-0.641	-0.128
Δ^2 LASM	0	45	-16.067[c]	-31.028[c]
LIMR	0	47	-1.432	-4.114
ΔLIMR	5	41	-0.666	-0.274
Δ^2 LIMR	0	45	-10.689[c]	-23.123[c]

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test. Maximum lag length was 9.
[e] Number of observations for ADF-GLS test. For the MZ_α test, $T = 45$.

Table 7. Results of ADF-GLS and MZ_α unit root tests with one structural break, 1950-1997, Canada

Variable	k [d]	T [e]	ADF-GLS	Tb[f]	MZ_α	Tb
LGDP	1	46	-3.781	1975	-24.632[a]	1975
ΔLGDP	1	45	-6.940[c]	1988		
LTHE	1	46	-3.332	1965	-18.978	1962
ΔLTHE	0	46	-8.144[c]	1960	-84.755[c]	1959
LPHE	1	46	-3.000	1971	-41.374[c]	1976
ΔLPHE	2	44	-4.543[b]	1961		
LHS	1	46	-4.582[b]	1985	-37.361[c]	1981
LASM	1	46	-3.427	1975	-24.330[a]	1976
ΔLASM	1	45	-5.887[c]	1975		
LIMR	4	43	-2.827	1990	-16.432	1965
ΔLIMR	3	43	-3.717	1987	-16.857	1987

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test using MAIC. Maximum lag length was 9.
[e] Number of observations for ADF-GLS test.
[f] Time of break.

Table 8. Results of LM unit root tests with one or two structural breaks, 1950-1997, Canada

Variable	k[d]	LM-1 test	Tb[e]	k	LM-2 test	Tb
LGDP	3	-3.690	1973	6	-5.494[a]	1970 1981[f]
ΔLGDP	1	-6.207	1976[f]			
LTHE	9	-10.257[c]	1974	9	-11.882[c]	1974 1992
LPHE	9	-4.020	1967	9	-9.093[c]	1967 1978
ΔLPHE	9	-2.775	1968			
LHS	9	-3.394	1963[f]	4	-5.054	1961 1980
ΔLHS	9	-3.453	1963	4	-4.392[b]	1967 1984
LASM	5	-4.236[a]	1977	4	-6.028[b]	1974 1987
LIMR	8	-4.646[b]	1977	6	-5.585[a]	1977 1992

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length. Maximum lag length was 9.
[e] Time of break.
[f] Break is not significant at the 10% level.

Table 9. Results of Causality Tests, 1950-1997, Canada

VAR Model 12A:			
LGDP → LASM	7.063 (0.0293)	LASM → LGDP	2.443 (0.2948)
LTHE → LASM	1.813 (0.4039)	LASM → LTHE	5.722 (0.0572)
LGDP → LTHE	1.104 (0.5757)	LTHE → LGDP	8.333 (0.0155)
VAR Model 12B:			
LGDP → LASM	6.234 (0.0443)	LASM → LGDP	3.534 (0.1708)
LPHE → LASM	2.055 (0.3576)	LASM → LPHE	3.775 (0.1515)
LGDP → LPHE	25.659 (0.0000)	LPHE → LGDP	6.010 (0.050)

Note: Arrows indicate the direction of causality under the alternative hypothesis; for all tests, the null hypothesis is that no relationship exists. P-values are in parentheses.

Table 10. Results of ADF-GLS and MZ_α unit root tests with no structural break, 1926-1999, Canada

Variable	k [d]	T [e]	ADF-GLS	MZ_α
LGNP	1	72	-3.566[b]	-23.802[c]
LIMR	4	69	-1.627	-6.744
ΔLIMR	8	64	-0.634	0.135
Δ^2 LIMR	0	71	-16.816[c]	-46.415[c]

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test. Maximum lag length was 11.
[e] Number of observations for ADF-GLS test. For the MZ_α test, $T = 71$.

Table 11. Results of ADF-GLS and MZ_α unit root tests with one structural break, 1926-1999, Canada

Variable	k [d]	T [e]	ADF-GLS	Tb[f]	MZ_α	Tb
LGNP	1	72	-4.447[b]	1981	-32.214[c]	1936
LIMR	4	69	-2.266	1937	-16.781	1937
ΔLIMR	1	71	-7.556[c]	1966	-42.158[c]	1970

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test using MAIC. Maximum lag length was 11.
[e] Number of observations for ADF-GLS test.
[f] Time of break.

Table 12. Results of LM unit root tests with one or two structural breaks, 1926-1999, Canada

Variable	k[d]	LM-1 test	Tb[e]	k	LM-2 test	Tb
LGNP	7	-6.043[c]	1949	7	-6.878[c]	1948 1989
LIMR	9	-5.064[b]	1977	9	-5.964[b]	1959 1977

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length. Maximum lag length was 11.
[e] Time of break.
[f] Break is not significant at the 10% level.

Table 13. Results of Causality Tests, 1926-1999, Canada

VAR Model:

LGNP → LIMR	22.912 (0.0004)	LIMR → LGNP	29.731 (0.0000)

Note: Arrows indicate the direction of causality under the alternative hypothesis; for all tests, the null hypothesis is that no relationship exists. P-values are in parentheses.

APPENDIX

A. Data sources

1. Data for *Age-standardized mortality rate per 100,000 population* come from Statistics Canada's *Health Indicators 1999* CD-ROM, catalogue 82-221-XCB.

2. The data source for the *Infant mortality rate* for 1921 and 1990 is the *Selected Mortality Statistics*, *Canada, 1921-1990*, Catalogue 82-548, while the data source for 1990 to 1999 is CANSIM, table 102-0030, available at http://cansim2.statcan.ca/cgi-win/cnsmcgi.exe?CANSIMFile=CII/CII_1_E.HTM&RootDir=CII/ .

3. *Perinatal mortality rate* data were taken from *Selected Infant Mortality and Related Statistics, Canada 1921-1990*, Catalogue 82-549 for 1921 to 1990, while the data source for 1990 to 1999 is Statistics Canada's *Health Indicators 1999* CD-ROM.

4. Because data for *Life expectancy* going back to 1920 are only presented for five-year intervals, life expectancy for Canada used in this article was computed using internal calculations on a yearly basis.[38] Calculations were made using data on death rates and population numbers provided by Statistics Canada's *Life Tables*.

5. *Total health expenditures* were constructed from the data series B513, *Total health expenditures,* of Statistics Canada's *Historical Statistics of Canada*[39] for 1945 to 1975 and from the series *Total health expenditures* in Table A.2.1 of CIHI (2001) for 1975 to 1999.

6. *Public health expenditures* were constructed from the following data sources:

a) Series H307, *All governments, gross general expenditure on health* from Statistics Canada's *Historical Statistics of Canada*, for 1965 to 1975
b) Series H150, *All governments, net general expenditure on health*, also from the *Historical Statistics of Canada*, for 1945 to 1969
c) Series *Total public sector health expenditure* from Table A.2.1 of CIHI (2001), for 1975 to 1999

7. *Total population* is the result of merging three series:

a) Series A1, from Statistics Canada's *Historical Statistics* , 1926 to 1961
b) Population series from the Statistics Canada's *Health Indicators 1999* CD-ROM, 1961 to 1970
c) Series V466668, Total population, from CANSIM, Table 051-0001, 1971 to 2000

---

[38] The authors would like to thank Allan Pollock, economist at Finance Canada, Economic Analysis and Forecasting Division, for computing these numbers.
[39] Leacy, F.H., ed. (1983) *Historical Statistics of Canada*. Ottawa: Statistics Canada, Catalogue 11-516-XIE. Available on the internet at http://www.statcan.ca:80/english/freepub/11-516-XIE/free.htm.

8. I*mplicit price indexes of government current expenditure* is constructed from series K176 of Statistics Canada's *Historical Statistics of Canada* (1945 to 1975) and the price index available in Table B.1 of CIHI (2001) (1975 to 1999). This price index was used to deflate both health expenditure series.

9. *GDP* data were taken from Finance Canada's economic forecasting model, CEFM. This GDP measure is equal to Statistics Canada's Series V1992067, *Gross Domestic Product at market prices*, from CANSIM, Table 380-0002.

8. *GNP* is constructed from the following data sources:

a) Series F13, *Gross National Product at market prices*, from Statistics Canada' *Historical Statistics of Canada*, 1926 to 1975
b) Series V499688, *Gross National Product at market prices*, from CANSIM, Table 380-0015, 1961 to 1999

9. GNP and GDP were deflated using a price deflator constructed from Series K172, *Implicit price indexes of gross national expenditure at market prices*, from Statistics Canada' *Historical Statistics of Canada* (1926 to 1975) and from Series V1997756, *Implicit price indexes, Gross Domestic Product at market prices* from CANSIM, Table 380-0003 (1961 to 1975).

10. OECD data:

a) From the *OECD Health Data 2002 CD-ROM*, 4th edition, version of August 20th 2002.
-Total expenditure on health per capita, national currency at 95 Total expenditure on health price level
-Infant mortality rate

b) From *OECD Economic Outlook*, no. 73, June 2003, to calculate GDP per capita national currency at 95 GDP price level:
-Nominal GDP in national currency
-GDP Deflator
-Total population

B. Results for Other OECD Countries

For purposes of comparison, the ADF-GLS, MZ_α, and LM unit root tests were applied to OECD data for Canada and five other OECD countries – Finland, Norway, Switzerland, the UK, and the USA – for the period 1960-1997. The results are summarized in Tables B1–B3. As discussed in section 4.2, the tests were applied to the levels, first differences, and second differences of the variables to determine the order of integration. The assumptions about constant and trend, and the choice of maximum lag length, were, in all cases, the same as for Canada.[40]

As can be seen from the table, the ADF-GLS test yields different results for different countries. Only one measure of health status, the infant mortality rate, was tested due to the lack of consistent data over the full period. This variable appears to be I(2) in Canada, Finland, Switzerland and the UK according to the ADF-GLS test, but I(1) in Norway and the USA. The MZ_α test indicates that LIMR is I(2) in every country. Real per capita health spending is I(1) for most countries using the ADF-GLS test, except for Finland and the USA where it is found to be I(2). LTHE is also I(2) for Canada, Finland, Norway and the USA using the MZ_α test. Finally, real per capita GDP appears to be I(2) with the ADF-GLS test in Finland, Norway, and Switzerland, while it is I(1) in Canada, the UK and the USA. The MZ_α test indicates that LGDP is I(2) for Canada, Finland and Norway and I(1) for the remaining countries. Thus the tests suggest different order of integration for the variables for all countries, except Finland. As was the case for the Canadian analysis, the I(2) results of the ADF-GLS and MZ_α tests for numerous countries suggest that applying unit root tests allowing for structural breaks would also be interesting for the OECD countries.

ADF-GLS and MZ_α tests results allowing for a structural break for the OECD countries are presented in Table B2. There is not much consistency between countries. As in Canada, LIMR is I(1) in Finland and the USA, but LIMR is I(2) in Norway and Switzerland, while being I(0) in the UK. LGDP was I(0) in Canada, same as Finland, the UK and the USA. However, LGDP is I(1) in Norway and Switzerland. Finally, LTHE is I(1) or I(2) for Canada depending on the test, while being I(1) in Norway and Switzerland. LTHE is I(0) in the UK and the USA, while being I(0) or I(1) in Finland.

Table B3 presents the results of the Lee and Strazicich (2003,2004) LM unit root test allowing for one or two structural breaks for the OECD countries. For these tests, the results are more consistent between countries. LIMR is I(0) in most countries using the LM-1 test, except for Switzerland where LIMR is I(2). Using the LM-2 test, LIMR is I(1) for Canada, Switzerland, the UK and the USA but is I(0) for the remaining countries. For the UK and the USA, one or both breaks are not significant at the 10% level of significance using the LM-2 test. LGDP is I(0) for every country, except Switzerland where LGDP is I(2) using the LM-1 test. LTHE is also I(0) for every country.

[40] Graphs of the levels and first differences of the OECD data are available from the authors upon request.

Table B1. Results of unit root tests, 1960-1997, OECD countries

Country	*LIMR*		*LGDP*		*LTHE*	
	ADF-GLS	MZ_α	ADF-GLS	MZ_α	ADF-GLS	MZ_α
Canada[d]	I(2)[c]	I(2)[c]	I(1)[c]	I(2)[c]	I(1)[a]	I(2)[c]
Finland	I(2)[c]	I(2)[c]	I(2)[c]	I(2)[c]	I(2)[c]	I(2)[c]
Norway	I(1)[a]	I(2)[c]	I(2)[c]	I(2)[c]	I(1)[a]	I(2)[c]
Switzerland	I(2)[c]	I(2)[c]	I(2)[c]	I(1)[a]	I(1)[c]	I(1)[b]
UK[e]	I(2)[c]	I(2)[c]	I(1)[c]	I(1)[c]	I(1)[c]	I(1)[c]
USA	I(1)[b]	I(2)[c]	I(1)[c]	I(1)[c]	I(2)[c]	I(2)[c]

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Data available between 1961 and 1997 for real per capita GDP.
[e] Data available between 1960 and 1996 for real per capita total health expenditures.

Table B2. Results of ADF-GLS and MZα unit root tests including a structural break, 1960-1997, OECD countries

Variable	k [d]	T [e]	ADF-GLS	T_b [f]	MZ_α	T_b
Canada						
LIMR	1	36	-3.053	1988	-11.999	1976
ΔLIMR	1	35	-6.238[c]	1983	-26.014[a]	1981
LGDP[g]	1	35	-4.043[a]	1974	-28.808[b]	1974
LTHE	1	36	-2.239	1978	-12.970	1967
ΔLTHE	1	35	-4.046[a]	1981	-15.527	1981
Finland						
LIMR	2	35	-2.367	1985	-11.639	1986
ΔLIMR	2	34	-4.389[b]	1982	-23.042[a]	1982
LGDP	1	36	-4.252[b]	1982	-42.217[c]	1984
LTHE	1	36	-2.478	1980	-250.294[c]	1967
ΔLTHE	1	35	-3.595[a]	1990	-14.826	1970
Norway						
LIMR	1	36	-3.114	1984	-13.897	1984
ΔLIMR	2	34	-2.634	1974	-6.286	1974
LGDP	1	36	-3.684	1975	-29.497[b]	1982
ΔLGDP	1	35	-4.450[b]	1987	-20.598	1987
LTHE	1	36	-2.342	1977	-15.697	1965
ΔLTHE	1	35	-4.500[b]	1974	-21.824[a]	1967
Switzerland						
LIMR	1	36	-2.962	1977	-13.227	1977
ΔLIMR	2	34	-3.553	1979	-10.963	1981
LGDP	1	36	-3.141	1975	-28.641[b]	1967
ΔLGDP	1	35	-5.203[c]	1979	-31.631[c]	1979
LTHE	1	36	-3.282	1974	-18.581	1974
ΔLTHE	1	35	-4.715[c]	1984	-23.524[b]	1984

Table B2 (continued). Results of ADF-GLS and MZα unit root tests including a structural break, 1960-1997, OECD Countries

Variable	k [d]	T [e]	ADF-GLS	T_b [f]	MZ_α	T_b
UK						
LIMR	1	36	-4.273[b]	1977	-33.963[c]	1977
LGDP	1	36	-4.755[c]	1976	-40.161[c]	1976
LTHE[h]	5	31	-14.44[c]	1973	-37.240[c]	1979
USA						
LIMR	1	36	-3.195	1976	-12.101	1973
ΔLIMR	1	35	-4.632[b]	1981	-143.815[c]	1979
LGDP	1	36	-4.925[c]	1965	-37.493[c]	1982
LTHE	1	36	-4.075[a]	1975	-46.876[c]	1966

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for ADF-GLS test using MAIC. Maximum lag length was 9.
[e] Number of observations for ADF-GLS test.
[f] Time of break.
[g] Data available between 1961 and 1997.
[h] Data available between 1960 and 1996.

Table B3. Results of LM unit root tests with one or two structural breaks, 1960-1997, OECD Countries

Variable	k[d]	LM-1 test	T_b[e]	k	LM-2 test	T_b
Canada						
LIMR	8	-4.701[b]	1977	6	-4.457	1977 1992[f]
ΔLIMR				0	-6.609[c]	1985 1987
LGDP[g]	6	-6.386[c]	1979	5	-7.548[c]	1980 1989
LTHE	6	-4.962[b]	1978	3	-6.708[c]	1973 1984
Finland						
LIMR	9	-5.252[c]	1982	7	-6.396[b]	1979 1990
LGDP	9	-4.484[a]	1980	3	-6.054[b]	1975 1989
LTHE	9	-5.496[c]	1977	5	-6.611[c]	1971 1983
Norway						
LIMR	7	-7.115[c]	1980	0	-6.894[c]	1979[f] 1988
LGDP	7	-7.320[c]	1986	5	-5.682[b]	1971[f] 1986
LTHE	9	-6.474[c]	1978	9	-5.758[b]	1978 1992
Switzerland						
LIMR	7	-3.795	1979	0	-5.184	1975 1988
ΔLIMR	4	-3.533	1991	0	-7.383[c]	1976[f] 1979
LGDP	9	-4.093	1979	9	-8.556[c]	1973 1987
ΔLGDP	1	-4.719	1982			

Table B3 (continued). Results of LM unit root tests with one or two structural breaks, 1960-1997, OECD

Variable	k[d]	LM-1 test	T_b[e]	k	LM-2 test	T_b
LTHE	9	-5.335[c]	1980	5	-6.546[c]	1974 1987
UK						
LIMR	9	-4.867[b]	1974	3	-5.302	1972[f] 1978
ΔLIMR				0	-5.168[c]	1980[f] 1985
LGDP	9	-4.843[b]	1978	5	-6.876[c]	1978 1986
LTHE[h]	9	-8.411[c]	1978	5	-15.277[c]	1973 1980
USA						
LIMR	8	-4.464[a]	1974	6	-4.157	1974 1984[f]
ΔLIMR				0	-4.504[b]	1973[f] 1982[f]
LGDP	7	-5.969[c]	1973	5	-5.774[b]	1972 1985
LTHE	5	-7.149[c]	1974[f]	5	-6.756[c]	1974[f] 1984[f]

[a] significant at the 10% level
[b] significant at the 5% level
[c] significant at the 1% level
[d] Lag length for LM test. Maximum lag length was 9.
[e] Time of break.
[f] Break is not significant at the 10% level.
[g] Data available between 1961 and 1997.
[h] Data available between 1960 and 1996.

C. Choice of lags for the unrestricted VARs

To determine the number of lags for the unrestricted VARs for Canada, a number of different criteria were used. First, the values of five information criteria were compared: the sequential modified LR test statistic (LR),[41] the Final prediction error (FPE), the Akaike information criterion (AIC), the Schwarz information criterion (SC) and the Hannan-Quinn information criterion (HQ). In the calculation of these criteria the maximum number of lags was set equal to six or seven for models in the 1960-1997 period, eight for the 1950-1997 period models and eleven lags for the 1926-1999 period models. Tables C1, C4 and C7 show the number of lags selected by each criterion.

Second, for each of the lag lengths shown in Tables C1, C4, and C7, joint tests of normality and heteroskedasticity were examined[42]. The results are presented in Tables C2, C5, and C8 for the three time periods. Third, tests for autocorrelation were done using the adjusted Q statistic and the LM test. These two statistics were generally calculated for up to 10 lags. The results appear in Tables C4, C6 and C9 for the three time periods.

The lag length actually used for the VAR models tests was generally the lag length selected by at least one information criterion and/or following the residuals who appeared to have most of these properties: normally distributed, homoskedastic, and not serially correlated.

[41] The sequential modified LR test statistics are at 5% level.

[42] When the lag length was very high, the results of the diagnostic tests always seemed to be worse. In those cases, the degrees of freedom were also very small. When this situation prevailed, diagnostic tests were done for short lag lengths only when they offered reasonable results.

Table C1. Results of diagnostic tests for the number of lags for the VARs with dummies, 1960-1997, Canada

Maximum number of lags	LR	FPE	AIC	SC	HQ
LASM, LGDP, LTHE (breaks in 1979, 1980 and 1982)					
Max=7	1	1	7	1	1
LASM, LGDP, LTHE (breaks in 1971, 1975, 1980, 1984, 1989, 1993)					
Max=6	1	2	6	1	2
LASM, LGDP, LPHE (breaks in 1976, 1979, and 1980)					
Max=7	2	2	7	2	7

Table C2. Results of tests on the residuals of the VARs, 1960-1997, Canada

Number of lags in the VAR	Jarque-Bera test of normality		Heteroskedasticity: White's test with squares of variables	
	Test statistic	P-value	Test statistic	P-value
LASM, LGDP, LTHE (breaks in 1979, 1980 and 1982)				
1	7.507	0.276	57.512	0.347
2	12.638	0.049	102.161	0.179
7	32.393	0	-	-
LASM, LGDP, LTHE (breaks in 1971, 1975, 1980, 1984, 1989, 1993)				
1	13.183	0.040	78.707	0.275
2	14.767	0.022	128.207	0.090
LASM, LGDP, LPHE (breaks in 1976, 1979, and 1980)				
2	6.674	0.352	88.465	0.526

Table C3. Results of residual Portmanteau tests for autocorrelation and of residual serial correlation LM tests, 1960-1997, Canada

Number of lags in the VAR	Adjusted Q-Statistic		LM Test	
	First sign of autocorrelation detected at lag…	P-value	First sign of autocorrelation detected at lag…	P-value
LASM, LGDP, LTHE (breaks in 1979, 1980 and 1982)				
1	2	0.016	1	0.047
2	3	0.028	-	-
7	8	0	1	0.092
LASM, LGDP, LTHE (breaks in 1971, 1975, 1980, 1984, 1989, 1993)				
1	2	0.008	1	0.028
2	3	0.061	-	-
LASM, LGDP, LPHE (breaks in 1976, 1979, and 1980)				
2	-	-	-	-

Table C4. Results of diagnostic tests for the number of lags for the VARs with dummies, 1950-1997, Canada

Maximum number of lags	LR	FPE	AIC	SC	HQ
LASM, LGDP, LTHE (breaks in 1960 (mean only), 1970, 1974, 1987, and 1992)					
Max=8	2	2	2	1	2
LASM, LGDP, LPHE (breaks in 1958, 1959, 1967, 1970, 1974, 1978, 1987)					
Max=8	2	2	6	2	2

Table C5. Results of tests on the residuals of the VARs with dummies, 1950-1997, Canada

Number of lags in the VAR	Jarque-Bera test of normality		Heteroskedasticity: White's test with squares of variables	
	Test statistic	P-value	Test statistic	P-value
LASM, LGDP, LTHE (breaks in 1960 (mean only), 1970, 1974, 1987, and 1992)				
1	7.732	0.258	65.701	0.487
2	10.859	0.093	96.200	0.643
LASM, LGDP, LPHE (breaks in 1958, 1959, 1967, 1970, 1974, 1978, 1987)				
2	14.557	0.024	105.010	0.564
6	27.418	0.0001	-	-

Table C6. Results of residual Portmanteau tests for autocorrelation and of residual serial correlation LM tests, 1950-1997, Canada

Number of lags in the VAR	Adjusted Q-Statistic		LM Test	
	First sign of autocorrelation detected at lag…	P-value	First sign of autocorrelation detected at lag…	P-value
LASM, LGDP, LTHE (breaks in 1960 (mean only), 1970, 1974, 1987, and 1992)				
1	2	0.0002	1	0.011
2	3	0.005	-	-
LASM, LGDP, LPHE (breaks in 1958, 1959, 1967, 1970, 1974, 1978, 1987)				
2	3	0.080	-	-
6	7	0	-	-

Table C7. Results of diagnostic tests for the number of lags for the VAR with dummies, 1926-1999, Canada

Maximum number of lags	LR	FPE	AIC	SC	HQ
LIMR, LGNP (breaks in 1948, 1959, 1977, and 1989)					
Max=11	8	11	11	2	5

Table C8. Results of tests on the residuals of the VARs, 1926-1999, Canada

Number of lags in the VAR	Jarque-Bera test of normality		Heteroskedasticity: White's test with squares of variables	
	Test statistic	P-value	Test statistic	P-value
LIMR, LGNP (breaks in 1948, 1959, 1977, and 1989)				
2	2.021	0.732	55.122	0.022
5	7.414	0.116	92.204	0.055
8	15.567	0.004	105.996	0.537
11	25.676	0	140.145	0.575

Table C9. Results of residual Portmanteau tests for autocorrelation and of residual serial correlation LM tests, 1926-1999, Canada

Number of lags in the VAR	Adjusted Q-Statistic		LM Test	
	First sign of autocorrelation detected at lag…	P-value	First sign of autocorrelation detected at lag…	P-value
LIMR, LGNP (breaks in 1948, 1959, 1977, and 1989)				
2	3	0.0003	1	0.0239
5	6	0.002	6	0.033
8	9	0	2	0.069
11	12	0	1	0.051