H.E. Frech III (USA), Stephen T. Parente (USA), Bianca K. Frogner (USA), John Hoff (USA)

# Comparing the sensitivity of models predicting health status: a critical look at an OECD Report on the efficiency of health systems

## Abstract

This article takes a critical look at a 2008 Report by the Organization for Economic Cooperation and Development (OECD), which attempted to identify the significant determinants that explain variation in health status across industrialized countries. The authors highlight the shortcomings of the health production model presented by the OECD Report such as the use of an incomplete measure of health status (i.e., life expectancy), the focus on the household production rather than individual demand, the partial measurement of health resource inputs, the choice of currency conversion factor, and the lack of rigorous econometric estimation procedures. The paper then compares how the OECD Report results compare to other estimates in the literature measuring health system efficiency. The choice of input and output variables, functional form, estimation methods, and time period lead to varying conclusions about the efficiency of the U.S. health system. Before concluding that the U.S. health system is the least efficient of the OECD countries, as suggested by the OECD Report, further work should be done to strengthen the OECD Report model.

**Keywords:** international comparison, health system, efficiency, determinants of health.

## Introduction

Researchers at the Organization for Economic Cooperation and Development (OECD) published a report in 2008 entitled, "Health Status Determinants: Lifestyle, Environment, Health Care Resources and Efficiency." The goal of the Report was to identify the significant determinants that explain variation in health status across OECD countries. The Report concludes that health care is highly productive in improving health outcomes and that productive efficiency varies greatly across countries. The Report goes further to provide country-specific estimates on productive efficiency. In this paper, we demonstrate that the country-specific conclusions are sensitive to the model specification. We present ways to improve the model, and also discuss how the data limits the ability to improve the model.

**Overview of health production.** According to Michael Grossman (1972, p. 223), "Health care is viewed as a durable capital stock that produces an output of healthy time." Health is the unobservable capital good that produces the healthy time that consumers actually value. Health depends on the preferences, values and choices of different individuals and different groups. Health is an aspect of human capital that is produced by household production. Health itself lasts into the future, and better health both enhances and extends one's life. Therefore, it is considered an element of human capital.

Health care is not directly valued independently of the health that it produces. Health care is one step away from the good that people actually value. The demand for health care is a derived demand, resulting from health care's productivity in producing health. Health insurance is removed by another step, since it depends on the productivity of health care in producing health and also on the consumer's subjective and idiosyncratic attitudes towards risk.

The OECD study takes the household production function approach to determining health status. In this approach, inputs are combined to produce the output, health. Estimation of health care productivity and the efficiency of different countries' health care systems requires adequate statistical controls for other determinants of health, to avoid confounding health care resource use or the efficiency of a country's system with other factors that operate in the society or the economy that are largely outside the health care system. Examples include healthy lifestyles, favorable cultures, high income, low pollution, good genes, favorable industrial and urban structure, and good education.

The most important distinguishing feature of *household* production is that productive decisions cannot be separated from the values and tastes of the consumers themselves. An important example is the time preference (impatience) of the consumer. Time preference has been shown to be related to a variety of health behaviors (Fuchs, 1980; Robb, Huston and Finke, 2008; Zhang and Rashad, 2008). Because of the intermingling of values and production, even with identical resources, different households, groups or countries make different choices and, therefore, end up with different health.

Health production data typically support diminishing returns (Baily and Garber, 1997; Frech and Miller, 1999; Fuchs, 2004; Garber and Skinner, 2008). The existence of diminishing returns implies that countries with heterogeneous populations (i.e., different consumers choosing different levels of health care) will appear, falsely, to be less efficient. In this context, Alan Garber and Jonathan Skinner (2008) point out that the U.S. is likely to be especially het-

erogeneous for two reasons. First, the health care insurance system is more varied. Second, regional variation in health care utilization (most of which cannot be explained by variation in health insurance) is more pronounced in the U.S. than in other rich countries.

One would expect diminishing returns in health production on both margins because of a tendency to allocate health care where it has the largest effect. On the extensive margin, one would expect the consumers with the most ability to benefit from the care to be the first ones to get care. Subsequent consumers to get care would be less likely to benefit. This type of rationing across consumers is called triage. On the intensive margin, the first type of care would be the most productive.

## 1. OECD's health production model

The OECD's health production model is as follows:

$$Y_{it} = \alpha_i + \beta HCR_{it} + \gamma \cdot SMOK_{it} + \phi \cdot DRINK_{it} + \\ + \theta \cdot DIET_{it} + \delta \cdot AIRPOL_{tt} + \sigma \cdot EDU_{it} + \lambda \cdot GDP_{it} + \varepsilon_{it}, \quad (1)$$

where output: Y = health status, variously measured. Inputs: *HCR* = health care resources *per capita*, measured two ways; *SMOK* = tobacco consumption in grams *per capita*; *DRINK* = alcohol consumption in litres *per capita*; *DIET* = consumption of fruit and vegetables *per capita* in kgs; *AIRPOL* = emissions of nitrogen oxide (NOx) *per capita* in kgs; *EDU* = share of the population (aged 25 to 64) with at least upper secondary education; *GDP* = GDP *per capita*, $\varepsilon$ = the error term, accounting for all omitted factors and randomness, $i$ = observations at the individual level, and $t$ = time.

**1.1. Output measures.** The health status measures for *Y* used in the OECD production function are life expectancy (LE) at birth for males, females and in total; LE at age 65 for males and females; Potential Years Life Lost (PYLL) for males, females and in total; and infant mortality. The Report stresses results on LE at birth. Health produces healthy time. Healthy time cannot exceed total time. Therefore, LE represents the maximum expected healthy time for an individual or a group. Indeed, an important line of theoretical work, closely associated with Isaac Ehrlich (1999), views the consumer as choosing a health level so as to optimally choose his life expectancy (LE). In principle, one might also adjust downward this measure to account for time in poor health. Further, one might try to directly measure healthy time. PYLL is one such measure that adjusts for certain external causes of death. Health adjusted life expectancy (HALE) is another measure that is used by other researchers, but was not included in the Report.

The authors of the report use a raw LE, rather than a morbidity-adjusted version because LE measures are available for more countries over more years, even if they are conceptually inferior. The reported correlations, for 2003 only, among the raw LE measures are fairly high, but correlations with PYLL, adjusted mortality and infant mortality are quite a bit lower (Joumard, André, Nicq and Chatal, 2008). As the Report notes, PYLL has an advantage over other measures, such as LE, in that it can be adjusted by cause of death to eliminate some of the causes of death that are due to other factors external to the health care system, such as accidents and violence. There is a natural way to do this using PYLL data given that the cause of death is recorded. One simply calculates a PYLL for deaths due to causes of death that are at least arguably sensitive to health care. One can also calculate PYLL for categories of diseases and analyze the effect of the health care system and other variables on PYLL by category, as is done in Miller and Frech (2004) for the respiratory, circulatory and cancer categories and in Or, Wang and Jamison for heart disease (2005).

One disadvantage that YPLL and LE at birth share is that both are contaminated by infant mortality; however, this contamination varies by cause of death. PYLL from cancer and heart disease are less contaminated by infant mortality than the general PYLL because infant deaths from these causes are fairly rare. PYLL by respiratory disease may be even more contaminated than LE at birth because respiratory disease is a major problem for infants. The Report adjusts PYLL to eliminate some external causes of death. It gives examples of excluded causes of death: land transport accidents, accidental falls, suicides and assaults, but it is not clear if this list is exhaustive (Joumard, André, Nicq and Chatal, 2008).

**1.2. Input measures.** There are two very different measures of health care input resources used in the Report. The first is a total spending variable that is aggregated over the entire health care system. One problem with this measure, as previous research suggests, is that the productivity differs for different types of care (e.g., spending on pharmaceuticals versus other spending, public versus private spending). The coefficient on the aggregate version captures a type of weighted average effect. A second problem is that the Report converts from spending in domestic currencies to a common currency using conversion rate, i.e., the Gross Domestic Product Purchasing Power Parity (GDP PPP) conversion factor, which may introduce a systematic error into the measure of health care inputs. GDP PPP is subject to its own fluctuations over time due to changing definitions and values of the basket of goods used to create the PPP.

The second measure of health care resources is an aggregated and partial measure of physical inputs. The Report creates an index of weighted health workers per 1,000 population, based on weighting a nurse as one half of a physician (Joumard, André, Nicq and Chatal, 2008). Written the same way, the OECD weight for physicians is 2.0 times the weight for nurses. The OECD numbers are expressed as the number of health workers per 1,000 population.

The other input measures include individual behaviors (i.e., smoking, drinking and eating habits), environmental factors (i.e., air pollution), and general economic factors (i.e., higher education and GDP). The choice of variables is based on support in the literature that suggests these factors influence health status, although directionality cannot be assumed. The choice in variables is also largely influenced by the availability of data. The variables also only provide a limited view on the complex set of factors that influence health. Also, none of these inputs take into consideration genetic or family history factors, which have been demonstrated to have significant effects on health and thus health care utilization.

**1.3. Empirical estimation.** The OECD Report's model uses partial fixed effects, with dummy variables entered only for countries, and not for years. This allows the constant term in the equation to only vary by each country. Time is not picked up by a year fixed effect or by a time trend. Time-invariant cross-sectional variation is absorbed into the country dummy variables. Since all the economic effects come from changes over time, this specification causes the estimated effects to be confounded with the passage of time. In the health care sector, the rapid pace of technological change and its impact on the productivity of health care is a concern, but this trend is not captured in the OECD Report's model.

The model is estimated by a Generalized Least Squares (GLS) method that corrects for heteroskedasticity (expected errors differing across observations) and serial correlation (errors being correlated over time). The correction for serial correlation is flexible, allowing for the serial correlation to differ among countries (Joumard, André, Nicq and Chatal, 2008). This correction for serial correlation may avoid the problem of spurious correlation that can overstate the relationship between variables that move together over time.

The basic regression is run for different output measures and, in some cases, separately for males and females. All continuous variables are in natural logs. Since the specification is log-log, all the coefficients can be interpreted as elasticities.

The issue of sample size is a bit confusing. The authors state at one point that the analysis is based on 23 countries from 1981 to 2003 (Joumard, André, Nicq and Chatal, 2008, p. 20), but then state in a footnote to that sentence that seven countries were excluded and that some countries' time series were not of full length because of data problems. This makes sense, since the largest reported sample size is 325. Complete data on 23 countries for 23 years would generate a sample size of 529. Also, some variables may have been interpolated.

## 2. Critiques of the OECD model

**2.1. Measurement.** The net effect is to bias upward the estimated apparent inefficiency of the U.S. health care system and probably to bias upward the estimated productivity of health care. Many factors that influence the production of health are either omitted or poorly measured. This confounds the true productivity of the input with other factors. The resulting coefficients include omitted variable bias. Because of inherent data limitations, this problem can only be minimized and not eliminated completely.

It is easy to estimate a statistical production relationship that is misleading. The estimates can either overstate or understate the true productivity of an input by confounding the true productivity of the input with other factors. Paradoxically, to avoid that confounding and, therefore, to estimate the productivity of one particular input, one must include all the other important inputs in the estimation process. For example, one might find a strong relationship between education and health if there were no other inputs in the model (i.e., in simple regression or inspecting a scatter plot of the data). But, education is closely related to other inputs, such as income, healthy lifestyle choices, certain types of culture and low pollution. The actual causation may be from these other variables, not education. The problem results from omitting one or more relevant, correlated variables from the analysis. Hence, it is called omitted variable bias.

The main problem is that important variables, especially lifestyle and cultural variables have been excluded. The model is already too truncated. The OECD approach implicitly counts all the variation at the country level as inefficiency in the health care system. This is a result of the interpretation given to the coefficient on the country-specific dummy variables and to the residual variation. In reality, this coefficient also picks up the effects of three other types of variation. First, as explained above, it picks up variation in excluded variables (e.g., lifestyle

variables) for which there is no data available. Second, it picks up the effects of systematically mismeasured variables, such as using GDP PPP exchange rates, rather than real health care PPP exchange rates. Third, it picks up random variation. This problem could be partially explored by augmenting the model with more relevant lifestyle variables, at the cost of fewer observations, but it cannot be explored by dropping variables from a model that is already incomplete.

**2.2. Statistical critiques.** Logging all the variables imposes a particular functional form on the data. This functional form is called a log-log, double-log or constant elasticity form. Using this log-log functional form, the estimated coefficients are elasticities, giving the percentage impact of a 1.0 percent increase in the variable. Thus, an estimate of 0.04 would imply that a doubling (i.e., a 100 percent increase) of health care resources would increase health status by 4 percent. This would be a large effect. The log-log form incorporates and imposes diminishing returns to the inputs. The log-log functional form exhibits diminishing returns if the estimated coefficients are less than 1.0 in absolute value. That is clearly the range of possible values here. The largest estimate of the effect of health care on any health measure, for example infant mortality, is -0.572. The largest estimate for any form of life expectancy is 0.061 (Joumard, André, Nicq and Chatal, 2008).

Technically, the primary emphasis is on econometric (panel regression) methods, rather than the operations research technique of data envelopment analysis (DEA). The panel method uses dummy variables for each country to control for all time-invariant differences across countries. These are called unit-specific fixed effects. There are no time fixed effects, so this is not a full fixed-effects approach. Further, there are no time trend variables. The Report interprets these estimated country-specific coefficients as the main part of the measure of health care efficiency, even though the coefficients pick up all fixed differences across countries, not just efficiency differences.

The DEA is generally inferior, less stable and less reliable. It relies on simply assuming that the apparently most efficient observations (highest observed output, lowest observed input) are on the efficient frontier (curve) and all others are inefficient. Efficiency by country is measured as the distance from the frontier to the actual data point. This method implicitly assumes that all unmeasured variation in health can be attributed to differences in health care system efficiency. This is the same implicit assumption that underlies the regression-based measures of health care system inefficiency. Recasting the analysis in a DEA framework does not make this assumption any more reasonable. The DEA approach is sensitive to measurement error, especially for observations at the extremes of the variables.

Further, the technique requires the use of a small number of inputs. Reportedly, results were not reasonable when several inputs were used (Joumard, André, Nicq and Chatal, 2008). This limitation exacerbates the problem of omitted variable bias. In the actual estimation, there were only three independent variables, health care resources, diet and a proxy for what the authors call economic, social and cultural status (ESCS) (Joumard, André, Nicq and Chatal, 2008). This last variable is taken from another data source, the OECD Programme for International Student Assessment (PISA). It is used here to stand in for both income and education to reduce the number of variables. The index is based on occupational status, parental education, family wealth, an index of home educational resources and an index related to culture in the home.

The Report's modeling does not include a variable for time. Thus, is it likely that some of the variables pick up the influence of time-related improvements in technology. Since resources devoted to health care have been increasing over time, confounding between health care resources and the passage of time are particularly a problem when measuring productivity of health care. The result will be to overstate the productivity of health care. Fuchs (2004) believes that this is a major problem, especially in time-series analyses such as the Report's. The effect should be smaller in cross-sectional analyses. In fact, it would vanish in cross-sectional analyses if technological diffusion were equal across countries. Further exacerbating the problem is that the results are likely to change based on the choice in time period of data collection; this choice poses a particular problem in that one cannot safely assume that technological improvements have the same impact on health in all time periods in all countries. Generally, the underlying challenge is that the Report does not adequately address technology and the literature provides few, if any, satisfactory measures to capture the impact of technology on health improvements beyond the inclusion of time variables.

There are two possible ways of dealing with the problem of not accounting for time trend. First, one could make the analysis a full fixed effects model, by adding a dummy variable for each year. Those year dummy variables would account for general exogenous shocks that affect all OECD countries, such as technological progress. That solution uses up a lot of degrees of freedom, hence statistical

power, because it requires the estimation of about 20 more coefficients. A partial solution that would be less costly in degrees of freedom would be to introduce a linear, or perhaps quadratic, time trend.

Another model, called the random effects model, is also commonly used for panel data. It is an adjustment for heteroskedasticity only – allowing the error term to differ by country and by year. Random effects models assume that there is no correlation between the country-specific effects and the explanatory variables (i.e., that the fixed effects, if any, are uncorrelated with the independent variables). Here, that seems clearly to be incorrect. Random effects coefficients are as vulnerable to omitted variable bias as the coefficients in an ordinary least squares. It is possible to use random effects and fixed effects in the same model, but that is rarely done.

The explanatory variables are contemporaneous with the health outcomes; there are no lags. This is a problem because it leads to measurement error and also the possibility of simultaneous equations bias. In terms of measuring the inputs into health production, using lags makes economic sense because it takes years for the effects of some variables, especially lifestyle ones, to take full effect. Not using lags will bias down the effects of observed and included lifestyle variables. Because the incorrect lag implies that the variable is not fully controlled for, it will introduce measurement error into the variable. This biases upwards the apparent inefficiency of the U.S. system, because the U.S. lifestyles are relatively unhealthy. Most of the prior literature uses lags. For example, in cross-sectional analysis, Comanor, Frech and Miller (2006), Miller and Frech (2004) and Zweifel and Ferrari (1992) use lags of about six to 10 years. In a panel of OECD data that is similar to what is used in the Report, Zweifel, Steinmann and Eugster (2005), test lags of differing lengths and report that a lag of 10 years seems to be the best. The only lag to be tested experimentally in the Report is on GDP (Joumard, André, Nicq and Chatal, 2008). The conceptual argument for lagging GDP is probably weaker than for many other variables.

The simultaneous equations problem arises because of possible reverse causation. A country may use many health care resources because its population is in poor health. That is, health outcomes may influence health care resources used, the reverse of what the Report's authors are trying to estimate. This effect would bias the apparent productivity of health care downward. While this is a new area of research, there is some evidence for this reverse causation in OECD data (Zweifel and Ferrari, 1992; Zweifel, Steinmann and Eugster, 2005). The use of lags would reduce concern about this issue. It is less likely that health outcomes in 2000 could have influenced health care spending ten years earlier in 1990 than that health outcomes could have influenced health care spending in the same year.

Another concern is spurious correlation, which is caused by what is called the unit root problem. The unit root problem is likely to be present in health care time series data (Miller and Frech, 2004). Asymptotically (as the sample size grows large), the serial correlation correction avoids the problem (Hamilton, 1994).

## 3. Comparison of alternative specifications of the OECD model

One can look to other estimates of country-specific apparent efficiency in the literature. These results are quite different from the ones stressed in the Report. We reproduce the results of the OECD model (Table 1) and compare them with alternative models (Tables 1, 2 and 3).

Table 1. Apparent U.S. efficiency differences: life expectancy using regression models

| Source | Measure of health care resources | Obesity controlled for? | Measure of health | Apparent inefficiency (relative to mean) |
|---|---|---|---|---|
| JANC, p. 25, Table 6 | Spending at health care PPP | No | LE at birth | -4.0 years |
| JANC, p. 34, Figure 9 | Spending at health care PPP | No | LE at birth | -2.5 years |
| JANC, p. 56, Figure A3.2 | Spending at health care PPP | No | Female LE at 65 | -0.5 years |
| JANC, p. 56, Figure A3.2 | Spending at health care PPP | No | Female LE at 65 | 0.0 years |
| CFM | Spending at health care PPP | No | Female LE at birth | -1.56 years |
| CFM | Spending at health care PPP | Yes | Female LE at birth | -0.53 years |
| CFM | Spending at health care PPP | No | Male LE at birth | -2.19 years |
| CFM | Spending at health care PPP | Yes | Female LE at birth | -1.56 years |
| CFM | Spending at health care PPP | No | Female LE at 60 | -1.46 years |
| CFM | Spending at health care PPP | Yes | Female LE at 60 | -1.00 years |
| CFM | Spending at health care PPP | No | Male LE at 60 | -0.62 years |
| CFM | Spending at health care PPP | Yes | Male LE at 60 | -0.18 years |

Source: Joumard, André, Nicq and Chtal (2008, pp. 25, 34, 56); Comanor, Frech and Miller (2006, p. 13).

Table 2. Country rankings by apparent efficiency: life expectancy

| JANC LE at birth health spending at GDP PPP | JANC female LE at birth physicians and nurses | JANC female LE at 65 health spending at GDP PPP | JANC female LE at 65 physicians and nurses | OWJ female LE at birth physicians | OWJ male LE at birth physicians | OWJ female LE at birth physicians | OWJ male LE at birth physi-cians |
|---|---|---|---|---|---|---|---|
| Iceland | Greece | Australia | Australia | Japan | Canada | Japan | Japan |
| Australia | Australia | France | France | Canada | Japan | Austria | Austria |
| New Zealand | Iceland | New Zealand | Canada | Australia | Australia | France | France |
| Korea | France | Canada | New Zealand | Austria | Austria | Australia | New Zealand |
| Greece | N. Zealand | Iceland | Iceland | Portugal | **U.S.** | Canada | U.K. |
| Canada | Korea | Korea | Korea | France | New Zealand | New Zealand | Portugal |
| Finland | Switzerland | Switzerland | Switzerland | New Zealand | U.K. | Belgium | France |
| Poland | Canada | Belgium | Greece | Germany | Portugal | Switzerland | Finland |
| Sweden | Sweden | Finland | Austria | Belgium | Finland | Portugal | **U.S.** |
| France | Netherlands | Sweden | Sweden | Greece | Germany | France | Belgium |
| Belgium | Germany | Austria | Finland | Switzerland | France | Germany | Germany |
| Ireland | Austria | Greece | **U.S.** | **U.S.** | Switzerland | Sweden | Greece |
| U.K. | Turkey | Poland | Germany | U.K. | Belgium | Greece | Switzerland |
| Czech Republic | Iceland | Netherlands | Poland | Finland | Greece | Spain | Canada |
| Netherlands | Finland | Norway | Netherlands | Ireland | Sweden | Italy | Spain |
| Switzerland | U.K. | Germany | Norway | Spain | Italy | Norway | Sweden |
| Austria | Poland | Ireland | U.K. | Sweden | Netherlands | **U.S.** | Italy |
| Germany | Czech Republic | **U.S.** | Ireland | Italy | Ireland | Netherlands | Ireland |
| Turkey | Denmark | U.K. | Denmark | Netherlands | Spain | Ireland | Netherlands |
| Norway | Norway | Denmark | Czech Republic | Norway | Norway | U.K. | Denmark |
| Denmark | **U.S.** | Czech Rep. | Hungary | Denmark | Denmark | Denmark | Norway |
| Hungary | Hungary | Hungary | Turkey | | | | |
| **U.S.** | | Turkey | | | | | |

Source: Joumard, André, Nicq and Chatal (2008, pp. 25, 34, 56); Or, Wang and Jamison (2005, pp. 543, 544).

Table 3. Rankings by apparent efficiency: PYLL heart disease and infant mortality

| OWJ female PYLL by heart disease physicians | OWJ male PYLL by heart disease physicians | JANC infant mortality health spending at GDP PPP | JANC infant mortality physicians and nurses | OWJ infant mortality physicians |
|---|---|---|---|---|
| Australia | **U.S.** | Czech Republic | Korea | Canada |
| Japan | Australia | Ireland | Czech Republic | Portugal |
| New Zealand | Canada | Finland | Greece | Austria |
| Finland | Finland | Korea | Iceland | Germany |
| Canada | Netherlands | Greece | Finland | Greece |
| Switzerland | New Zealand | Poland | Poland | U.K. |
| **U.S.** | Denmark | New Zealand | France | Australia |
| Sweden | Switzerland | Australia | New Zealand | France |
| Netherlands | Belgium | Hungary | Australia | **U.S.** |
| Denmark | U.K. | Sweden | Hungary | New Zealand |
| Belgium | Sweden | Belgium | Denmark | Japan |
| France | Japan | France | Sweden | Switzerland |
| Portugal | France | Denmark | Germany | Denmark |
| U.K. | Portugal | Ireland | Austria | Italy |
| Italy | Italy | Canada | U.K. | Spain |
| Spain | Germany | U.K. | Canada | Finland |
| Greece | Norway | Austria | Netherlands | Belgium |
| Germany | Austria | Germany | Ireland | Sweden |
| Austria | Spain | Netherlands | Norway | Norway |
| Ireland | Greece | Norway | Switzerland | Ireland |
| Norway | Ireland | Switzerland | Turkey | Netherlands |
| | | **U.S.** | **U.S.** | |
| | | Turkey | | |

Sources: Joumard, André, Nicq and Chatal (2008, p. 69); Or, Wang and Jamison (2005, pp. 545, 546).

In the Report, it is more convenient to refer to the measures as efficiency measures, bearing in mind that they can take on a value that is either positive or negative. The estimates seem implausibly large. For example, in using expenditures and looking at LE at birth, the U.S. country-specific efficiency measure is -4.0 years. Other apparently low performers are Hungary at -3.1 and Denmark at -1.5 years. The U.S. comes out as the worst of all and quite a bit worse than the next rich country (Denmark). At the other extreme, for Iceland the score is 2.6 and for Australia, it is 2.5 years (Joumard, André, Nicq and Chatal, 2008). This means that, if the U.S. had as an efficient a health care system as Iceland's, U.S. life expectancies would be greater than they are by 6.6 years. If it was as efficient as Australia's, LE would be 6.5 years greater (Joumard, André, Nicq and Chatal, 2008).

The OECD results are not robust to using different measures of health care resources or to using different measures of health outcome. The OECD Report states that country efficiency rankings are roughly similar using LE at 65 versus at birth (Joumard, André, Nicq and Chatal, 2008). But that's apparently not so for the ranking of the U.S. As mentioned above, the U.S. inefficiency is -4.0 years for LE at birth (Joumard, André, Nicq and Chatal, 2008). But, for female LE at 65 (the only one presented) the U.S. estimated inefficiency using health spending is about -0.5 years. The U.S. rank is 17th of 23 and the U.S. now does better than the U.K. and Ireland (Joumard, André, Nicq and Chatal, 2008). Returning to LE at birth, when health care resources are measured in physical terms, the U.S. inefficiency estimates drops to -2.5 years and it ranks only above Hungary (Joumard, André, Nicq and Chatal, 2008). The difference between these two probably partly reflects the bias in using GDP PPP exchange rates for converting health spending to dollars.

Looking at female LE at 65, for the physical health resources measure, the U.S. does even better, with an inefficiency of about zero and a ranking at the median, 12th out of 23. See Table 1 for a display of the different apparent inefficiency estimates from the various regression versions in the Report and in Comanor, Frech and Miller (2006). As one can see, the measure of relative inefficiency varies greatly, depending on how the inputs are measured, on whether obesity is controlled for and also on which measure of LE is used.

These estimates in the Report are larger than those in the literature. For example, in a comparable work for a single cross section, Comanor, Frech and Miller (2006) estimated the relative shortfall, for LE at birth, of the U.S. at -1.56 years for males and -0.53 for females when obesity was controlled for. This is based on the residual for the U.S. One would get an identical answer by inserting a dummy variable for the U.S. But, a full set of country dummies cannot be used in a cross section because that would lead to negative degrees of freedom – mathematically impossible to estimate. This is far smaller of an effect (less in absolute value) than the -4.0 years from the Report. Even when they do not control for obesity, their estimated shortfall is -2.19 years for males and -1.56 for females (Comanor, Frech and Miller, 2006), substantially smaller than the shortfall estimated in the Report.

**3.1. Or's model.** In highly related work, also using a panel of OECD countries, Or (2000a, 2000b) also found health care resources to be highly productive. The earlier work focused on the effects of health spending on PYLL and will be discussed below. The later work used very similar techniques to the current Report, but measured health resources with the physician/population ratio only. Or found estimated elasticities of about 0.10 for LE at 65, roughly twice as large as the effects found in the Report.

The effects of health care resources on PYLL are larger in terms of elasticities, ranging from -0.062 for males to -0.089 for females and -0.072 to the total sample. However, they are not very precisely estimated, being statistically insignificant for males and statistically significant at only the 5 percent level for females. In Or's previous work with similar panel data, she found less consistency, but even larger effects, as high as -0.38 for women and -0.28 for men (2000b). The results are not exactly comparable because the PYLL in the Report has apparently been defined to exclude deaths from land transport accidents, accidental falls, suicides and assaults (Joumard, André, Nicq and Chatal, 2008).

As is discussed above, the effects of any input on LE and PYLL are not comparable, even though they both are expressed in years. These large elasticities for effects on PYLL do not translate into large effects on life years. For example, if we take the Report's largest estimate of -0.072 as if it were correct, this implies that a doubling of resources would cause a 7.2 percent decline in PYLL. But, the average is only 3,158 per 100,000 people 0-70 years old, or 0.032 years per person per year. Reducing that by 7.2 percent, we get an increase of 0.0023 years per year. Even accumulating these effects over 70 years, this is only 0.15 years. The effect of health care on PYLL is small because health care does not have so much effect at the earlier ages. Even though the effect is statistically significant, it is not so significant from a scientific or policy viewpoint. In the Report, as we have seen, even poorly measured lifestyle and socioeconomic factors seem to be much more power-

ful for PYLL than for LE. It is not clear why this should be, especially with the exclusions of some lifestyle-related causes of death.

**3.2. Or, Wang, Jamison's model.** More recently, Or, Wang and Jamison (2005) analyzed the impact of the physician/population ratio on LE in a panel of OECD countries. In their analysis that is most comparable to the Report, Or, Wang and Jamison (2005) find elasticities with respect to physicians varying between 0.037 and 0.077 for the LE analyses. This is a substantially bigger effect than the Report finds. Note that even Or, Wang and Jamison's (2005) lowest elasticity indicates that health care is very productive. A doubling of resources would raise female LE at birth by over two years. In percentage terms, LE at age 65 is substantially more sensitive to health resources than at earlier ages, roughly twice as sensitive. Most of the difference is simply a mechanical implication of the fact that most death occur after age 65. LE at birth is about 70 to 75 years in most of this data, while LE at 65 is about 15 to 20 years. Thus, an elasticity at the later years of twice as high implies, somewhat paradoxically, a smaller number of life years gained, not a larger number of life years gained.

The U.S. comes out generally much higher in Or, Wang and Jamison's (2005) model for LE at birth for females (Table 2). The U.S. ranks 12th out of 21 in health care efficiency, ranking higher than the U.K. Norway, and Sweden. In their rankings for males, the U.S. is above the mean and the median, ranking 5th out of 21. The rankings are not consistent across measures of health. In infant mortality, the U.S. is 9th of 21. Looking at LE at 65, the U.S. is 17th of 21 for females and 9th of 21 for males. In avoiding premature mortality from cardiovascular disease, the U.S. health care system is superior, ranked 7th of 21 for females and first, the top performer, for males (Or, Wang and Jamison, 2005).

Or, Wang and Jamison's estimates are not directly quantitatively comparable to the Report's. One problem is that the Report did not look at LE at 65. Also, Or, Wang and Jamison (2005) use a more flexible alternative that allows the (slope) coefficient on health care to vary across countries, as well as the constant. In effect, they use a dummy variable for each country to control for unexplained fixed effects, like the Report. But, they do not interpret the coefficient on this dummy variable as measuring the efficiency of the health care system. Rather, they allow for the effect of health care resources (here, the physician/population ratio) to vary across countries. They interpret differences in the coefficient on this variable as the efficiency difference across countries. Their efficiency measures are, therefore, differences in slopes across countries, while the OCED Report's efficiency differences are differences in the constant term across countries. Estimating efficiency by differences in slopes is conceptually superior to the Report's interpretation. It is less confounded by other influences on health. Still, the Or, Wang and Jamison's approach is vulnerable to a weaker version of same the criticism, i.e. the slope of the production function can also differ across countries because of confounding influences (Garber and Skinner, 2008).

In the Report's estimates for infant mortality, the U.S. comes out poorly, either the lowest or the second lowest (second to Turkey) (Joumard, André, Nicq and Chatal, 2008). This is the health measure that seems to be most sensitive to omitted variables, especially those reflecting cultural and lifestyle influences. But, the results for infant mortality are quite different in Or, Wang and Jamison (2005), where the U.S. comes out in the middle of the pack (Table 3). One of the main reasons for this is that Or, Wang and Jamison do not attribute estimated country differences to health care productivity, while the OECD Report does.

**3.3. Pauly's model.** The Report states that the weighted health workers is *ad hoc*, but a weighting of this sort can be based on objective market data, as was done by Mark Pauly (1993). Pauly simply computes the percentage of the population and the workforce who work in health care (1993, p. 156). Pauly includes a much broader array of workers (including many unskilled and semiskilled workers) and uses relative wages in the U.S. to form the weights. Thus, it is conceptually superior to the more limited measure. Further, the difference is quantitatively important. Physicians and nurses in total make up only 18.6 percent of the U.S. health care workforce, 3.4 percent for physicians and 15.2 percent for nurses (Bureau of Labor Statistics, 2008). The weight for physicians is 4.83 times the weight for other workers.

Using Pauly's physical input measure, the most comprehensive, the U.S. resource use is 6th of 12, slightly below the mean. Using the more narrow measure based on weighted physicians and nurses only, the U.S. is 4th out of 14. Looking only at physicians, the U.S. is only 9th of 18, and again, slightly below the mean. Clearly, the U.S. uses relatively more nurses and less of other types of nonphysician workers than the other OECD countries, so the Report's measure overstates U.S. resource use. Most importantly, the U.S. is not a high user of labor resources in its health care system. Using the GDP PPP exchange rates to calculate real resource use is highly misleading. Using that data in a health production model creates a large bias towards inaccurately portraying the U.S. system as inefficient in

producing health with health care resources. Another point to note is that these physical measures differ quite a lot, even though they are based on health care personnel. The creation of broader indexes with actual weights, in the spirit of Pauly's work, would be welcome.

Another approach is to concentrate on physicians. One could also examine the number of physician visits *per capita*. On this measure, the U.S. is quite low, at about 3.6, while Germany is 8.5 and France is 7.0. The U.S. ranks 15th out of 18 (Van Doorslaer, Masseria and Koolman, 2006). One could also simply use the number of physicians per 1,000 population. This is one of the measures of health resources (as opposed to spending) used by Anderson, Reinhardt, Hussey and Petrosyan (2003), and Anderson, Frogner and Reinhardt (2007) in their descriptive analysis.

It is also the only approach of Or, Wang and Jamison in their health production study (2005). Returning to the Report and looking at infant mortality, weighted physicians and nurses have a notable effect, with an elasticity of -0.440. This is a large effect, though not as large at GDP *per capita*, education or alcohol consumption. Or, Wang and Jamison (2005) also estimated large effects of physicians alone, -0.548 in the most comparable formulation. Considering the importance of omitted lifestyle variables and the measurement problems with infant mortality, discussed above, it is difficult to know what weight to give to the results for infant mortality.

**3.4. Health expenditure model.** The OECD researchers also used a measure of resources in monetary terms. As is discussed above, they converted the spending in any one country and year to U.S. dollars at constant prices, using the GDP PPP. Since the U.S. has higher relative health care prices and wages than other countries, the use of the GDP PPP systematically overstates U.S. health care resources used, thus understates U.S. health care productivity. This is one of the reasons for using the partial physical input measures discussed above (also see Frech, 2009). The results show a generally larger effect on health outcome than did weighted physicians and nurses. Also, the elasticities are more constant across differing LE measures. The elasticities for LE range from 0.035 to 0.061. These results are roughly comparable to, though slightly larger than, the results for the productivity of pharmaceuticals in cross sectional analyses (Frech and Miller, 1999; Miller and Frech, 2004; Shaw, Horrace and Vogel, 2005). These works do not contain reliable estimates for non-pharmaceutical health care, probably because it is so correlated with income. Nixon and Ulmann (2006) obtained a lower estimate, roughly half of the Report, using the same GDP PPP exchange rate as the Report. These pharmaceutical productivity papers used the health care PPP exchange rate, while the Report used the GDP PPP exchange rate. Zweifel, Steinmann and Eugster (2005) use a quadratic function, not a log-log function. Therefore, elasticities are not constant and must be calculated at some specified values. When calculated at the mean, these values are 0.035 for females and 0.045 for males (2005, pp. 135-137). Their equation has fewer controls on lifestyle and socioeconomic variables and does not include country-specific fixed effects. One might therefore have expected a larger effect for health care spending, rather than the somewhat smaller effect they obtain.

Using health expenditures, the estimates rise very little with age (comparing LE at birth to LE at age 65). For females, the elasticity is 0.035, rising to only 0.051 at 65. For males, the corresponding elasticities are 0.045 and 0.061. In comparable work, also using a panel of OECD countries (for a slightly earlier period, 1970-2000) for LE at 60, Zweifel, Steinmann and Eugster (2005) found elasticities that are somewhat lower. Comparing these Report's results for LE at 65 to those for LE at birth implies that health care has a substantially smaller effect on life years for older people.

For PYLL, the effect of health care resources is much larger than it is for LE. The elasticities vary in a tight range -0.272 to -0.300. Also, all of these results were highly statistically significant, in contrast to the PYLL estimates using weighted physician and nurses. Some of Or's previous work also used expenditures. She found less consistency across males and females, and also smaller estimates, -0.18 for women and a very small, -0.04 for men (2000a). Turning to infant mortality, again, the estimated effects are large, at -0.572. This is a large effect and larger than any single other input. Nixon and Ulmann (2006), find similarly large effects of health care spending on infant mortality (2006). As with the other infant mortality estimates, it is hard to know how to interpret the results, given the specification and data problems discussed above.

Comparing the results with the two measures, one can say that the real expenditures measure, even based on the inappropriate GDP PPP exchange rate, gives generally larger effects on health measures. However, both measures are flawed, as is explained above. Also, there are major problems of omitted variables that are correlated with both measures. So, the estimates confound the influ-

ences of the omitted variables and health care resources. Which flawed measure is preferable for estimating the effect of health care resources on health is not clear.

## Conclusion

A close look at the results for the productivity of health care from these alternative approaches suggests that most estimated coefficients are broadly stable in level and significance. There are some deviations, for example, the estimated spending elasticity is also somewhat higher in models without GDP, reflecting the correlation between GDP and health spending. When health care resources are measured by the number of practitioners, estimations are less stable across alternative scenarios.

On the other hand, the results for country-specific relative efficiency are fragile. They differ greatly across different approaches. In particular, the estimates for U.S. relative efficiency and rank in efficiency differ greatly across approaches.

## References


1. Anderson, G.F., Reinhardt, U.E., Hussey, P.S., Petrosyan, V. (2003). It's the Prices, Stupid: Why the United States is So Different from Other Countries, *Health Affairs*, Vol. 22, No. 3, pp. 89-105.
2. Anderson, G.F., Frogner, B.K., Reinhardt, U.E. (2007). Health Spending in OECD Countries: An Update, *Health Affairs*, Vol. 26, No. 5, pp. 1481-1489.
3. Baily, M.N., Garber, A.M. (1997). Health Care Productivity, *Brookings Papers on Economic Activity: Microeconomics*, pp. 143-215.
4. Comanor, W.S., Frech, H.E. III, Miller, Jr., R.D. (2006). Is the United States an Outlier in Health Care and Health Outcomes? A Preliminary Analysis, *International Journal of Health Care Finance and Economics*, Vol. 6, No. 1, pp. 3-23.
5. Frech, H.E. III (2009). The OECD's Study on Health Status Determinant: Roles of Lifestyle, Environment, Health-Care Resources and Spending Efficiency: An Analysis, AEI Working Paper #145.
6. Frech, H.E. III, Miller, Jr., R.D. (1999). *The Productivity of Health Care and Pharmaceuticals: An International Comparison*, Washington, D.C.: AEI Press.
7. Frech, H.E. III, Mobley, L.R. (2000). Efficiency, Growth, and Concentration: An Empirical Analysis of Hospital Markets, *Economic Inquiry*, Vol. 38, No. 3, pp. 369-384.
8. Fuchs, V.R. (1980). Time Preference and Health: An Exploratory Study, in *Economic Aspects of Health*, Fuchs, V.R. (Ed.), National Bureau of Economic Research, University of Chicago Press.
9. Fuchs, V.R (2004). More Variation in Use of Care, More Flat-Of-The-Curve Medicine, *Health Affairs*, doi:10.1377/hlthaff.var.104.
10. Garber, A.M., Skinner, J. (2008). Is American Health Care Uniquely Inefficient? *Journal of Economic Perspectives*, Vol. 22, No. 4, pp. 27-50.
11. Greene, W. (2004). Distinguishing between Heterogeneity and Inefficiency: Stochastic Frontier Analysis of the World Health Organization's Panel Data on National Health Care Systems, *Health Economics*, Vol. 13, No. 10, pp. 959-980.
12. Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health, *Journal of Political Economy*, Vol. 80, No. 2, pp. 223-255.
13. Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press.
14. Joumard, I., Andre, C., Nicq, C., Chatal, O. (2008). Health Status Determinants: Lifestyle, Environment, Health Care Resources and Efficiency, OECD Economics Department Working Papers No. 627.
15. Miller, R.D., Jr., Frech, H.E. III (2004). *Health Care Matters: Pharmaceuticals, Obesity and the Quality of Life,* Washington, D.C.: AEI Press.
16. Nixon, J., Ulmann, P. (2006). The Relationship between Health Care Expenditure and Health Outcomes: Evidence and Caveats for a Causal Link, *European Journal of Health Economics*, Vol. 7, pp. 7-18.
17. Or, Z. (2000a). Determinants of Health Outcomes in Industrialized Countries: A Pooled, Cross-Country, Time-Series Analysis, OECD Economic Studies No. 30 2000/1. Available at: http://www.oecd.org/eco/growth/2732311.pdf [Accessed 2013 February 21].
18. Or, Z. (2000b). Exploring the Effects of Health Care on Mortality across OECD Countries. OECD Labor Market and Social Policy Occasional Papers, No. 46. Available at: http://dx.doi.org/10.1787/716472585704 [Accessed 2013 February 21].
19. Or, Z., Wang, J., Jamison, D. (2005). International Differences in the Impact of Doctors on Health: A Multilevel Analysis of OECD Countries, *Journal of Health Economics*, Vol. 24, pp. 531-560.
20. Pauly, M.V. (1993). U.S. Health Care Costs: The Untold True Story, *Health Affairs*, Vol. 12, No. 3, pp. 152-159.
21. Robb, C.A., Huston, S.J., Finke, M.S. (2008). The Mitigating Influence of Time Preference on the Relation between Smoking and BMI Scores, *International Journal of Obesity*, Vol. 32, No. 2, pp. 1-8.
22. Schmidt, P., Sickles, R.C. (1984). Production Frontiers and Panel Data, *Journal of Business & Economic Statistics*, Vol. 2, No. 4, pp. 367-374.
23. Van Doorslaer, E., Masseria, C., Koolman, X. (2006). Inequalities in Access to Medical Care by Income in Developed Countries, *Canadian Medical Association Journal*, Vol. 174, No. 2, pp. 177-183.

24. Zhang, L., Rashad, I. (2008). Obesity and Time Preference: The Health Consequences of Discounting the Future, *Journal of Biosocial Science*, Vol. 40, No. 1, pp. 97-113.
25. Zweifel, P., Ferrari, M. (1992). Is There a Sisyphus Syndrome in Health Care?" in *Health Economics Worldwide*, Zweifel, P., Frech, H.E. III (Eds.), Amsterdam: Kluwer, pp. 311-330.
26. Zweifel, P., Steinmann, L., Eugster, P. (2005). The Sisyphus Syndrome in Health Revisited, *International Journal of Health Care Finance and Economics*, Vol. 5, No. 2, pp. 127-145.