Social Science & Medicine 93 (2013) 21–28

Contents lists available at SciVerse ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Assessing health system performance: A model-based approach

John Gerring[a,*], Strom C. Thacker[a,b], Ruben Enikolopov[c], Julián Arévalo[d], Matthew Maguire[a]

[a] Department of Political Science, Boston University, 232 Bay State Road, Boston, MA 02215, United States
[b] Department of International Relations, Boston University, 152 Bay State Road, Boston, MA 02215, United States
[c] New Economic School, Nakhimovsky pr. 47, Moscow, 117418, Russia
[d] Facultad de Economía, Universidad Externado de Colombia, Calle 12 No. 1-17 Este, Bogotá, Colombia

ARTICLE INFO

Article history:
Available online 14 June 2013

Keywords:
Global health
Health policy performance
Health policy assessment

ABSTRACT

It is difficult to assess countries' relative success in addressing issues of public health because countries are subject to very different background conditions. To address this problem we offer a model-based approach for assessing health system performance. Specifically, an index of public health is regressed against a vector of variables intended to capture economic, educational, cultural, geographic, and epidemiological endowments. The residual from this model is regarded as a plausible measure of public health performance at the national level.

We argue that a model-based approach to performance is informative for policymakers and academics as it focuses attention on those aspects of a country's health profile that are not constrained by structural factors. This sharpens comparisons across countries and through time, and also allows one to evaluate the degree to which health systems have lived up to their potential.

Introduction

How should one assess country performance in the health sector? This is a critical issue for publics, policymakers, and policy specialists. Without reliable indicators of performance one cannot gauge the success or failure of private and public efforts. "Improvement" and "deterioration" become matters of speculation, and accountability for policy choices is impossible to establish.

Here, we limit ourselves to a discussion of health sector effectiveness, leaving aside issues of equity and efficiency. Even so, one is at pains to evaluate performance across countries and through time. The problem is not an absence of indicators; there are a plethora of indicators of health performance to choose from (Backman et al., 2008; Kruk & Freedman, 2008; Murray & Evans, 2003). However, most policy-related indicators (e.g., immunization rates) are partial; they do not provide a total picture of health sector performance in a country. Moreover, few indicators provide sufficient country or historical coverage to judge performance on a global scale and over time.

Mortality-based outcome measures such as infant mortality or life expectancy compensate for these shortcomings. However, they are strongly affected by factors that lie outside the health sector. Indeed, a large portion of the variance in health outcomes may be explained as a correlate of economics, geography, education, or disease vectors that affect countries differently for reasons having nothing to do with the health sector in a particular country. While it may be meaningful to view life expectancy as a measure of the performance of health sectors across similarly situated countries such as the United States and Canada, it is virtually meaningless to compare this statistic across countries with vastly different endowments. We do not learn much, if anything, about the relative success of health sectors in the US and Sri Lanka by comparing life expectancy in these two countries—unless, that is, we can find a way to partial out the causal effect of background factors.

This is the intuition behind most international comparisons. When writers point to the extraordinary achievements of countries like Costa Rica, Cuba, and Sri Lanka they are (implicitly) comparing human development achievements in these countries relative to certain baseline characteristics thought to lie outside—or at least be separable from—the social policy sector (Caldwell, 1986; Ghai, 2000; Halstead, Walsh, & Warren, 1985; McGuire, 2010; Mehrotra & Jolly, 1997; Riley, 2007). Unfortunately, this handicapping exercise is rarely conducted in a systematic and explicit fashion. One

* Corresponding author.
E-mail address: jgerring@bu.edu (J. Gerring).

0277-9536/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.socscimed.2013.06.002

may limit the comparison to other developing countries. But this means that we can say nothing about the relative success of the health sectors in countries with very different levels of development. Moreover, it requires a categorical judgment about a quintessentially scalar phenomenon—development. Separating "developing" and "developed" countries is highly arbitrary at the breakpoint. Finally, it excludes other factors that also affect public health performance but are not the responsibility of the health sector.

To overcome these difficulties, model-based approaches have been developed for use in health policy (e.g., Wang, Jamison, Bos, Preker, & Peabody, 1999; WHO, 2000 [reviewed by Jamison & Sandbu, 2001]), human development (e.g., Kakwani, 1993), and other policy contexts (e.g., Ndulu & O'Connell, 2007). The approach laid out in this study differs from previous studies in four respects: (a) the spatial and temporal breadth of the dataset (including all country-years from 1960 to 2010), (b) the inclusion of an extensive set background factors, (c) close attention to problems of specification (including a large number of robustness tests), and (d) both cross-sectional and fixed-effect models.

In the first section of the paper we define a strategy for measuring health outcomes and show changes over time in the global distribution of public health. In the second section, we construct a cross-sectional model of health sector performance that includes background factors with expected impact on a country's health. In the third section, we use the residuals, or unexplained variance, from these models to assess countries' relative achievement and improvement across the 1960–2010 period. In the fourth section, we construct a fixed-effect model whose purpose is to track within-country change over time. In the fifth section, we conduct a series of sensitivity tests in order to probe the robustness of the results from these models. The concluding section reflects on the interpretation and possible uses of model-based measures of policy performance. Note that *country-years* are the relevant units of analysis in the following discussion.

Measuring public health

To measure the health of societies we focus on mortality data, specifically *life expectancy* and the *infant mortality rate* (IMR), understood as the number of babies that do not survive to age one per 1000 live births, transformed by the natural logarithm (to account for expected non-linearities). Of all possible health indicators, these are probably the most reliable and the most widely available through time and across countries.

While child mortality (deaths before the age of five) is sometimes regarded as a more valid statistic it is so highly correlated with IMR that the differences are of very little practical significance. We opt for IMR solely by reason of its more extensive coverage. With respect to life expectancy, one might prefer a more sensitive measure that takes account of varying levels of morbidity such as disability-adjusted life years (DALY) or health-adjusted life expectancy (HALE). Unfortunately, these adjustments are possible only in recent years and tend in any case to be highly correlated with (unadjusted) life expectancy, as we shall see.

Although IMR (ln) is a component of life expectancy, our index combines both statistics in a single indicator of public health, understood as the sum of the (standardized) values for the two component variables. We adopt this approach for several reasons. First, by incorporating data from two sources we are able to build a larger sample of observations, one that is also probably more typical of the total population of nation-states that we seek to represent. The two statistics are highly correlated, so this statistical manipulation imposes little loss of information. Second, these two mortality-based statistics describe somewhat different components of the topic. Although life expectancy is the "summary" concept, it might be argued that loss of life at a very early age is a greater human tragedy since it represents the loss of nearly a whole life. Finally, because of the greater vulnerability of newborns, IMR tends to be sensitive to policy interventions and societal behavioral changes to a much greater degree than life expectancy, as evidenced by the greater variance of IMR. For all these reasons, the combination of life expectancy and IMR (ln) offers a more reliable, more sensitive, and more insightful measure of public health than either would provide on its own. Data sources are explained in Table A1.

Readers may be curious to know how the resulting index compares with other measures of public health such as health-adjusted life expectancy (HALE), disability-adjusted life years (DALYs), child mortality, malnutrition (as proxied by height for age), as well as broader quality-of-life measures such as the poverty headcount ratio (percentage of the population living on less than $2 a day), and the UNDP's Human Development Index. Table 1 shows inter-correlations among these alternate measures and (a) the components of our index and (b) the composite index itself. Not surprisingly, our public health index is highly correlated with other measures of public health and with other quality-of-life measures. Good (bad) things generally go together.

One of the benefits of an index with broad coverage is that one can employ it to compare global distributions at varying points in time. Fig. 1 displays a kernel density plot of the distribution of public health in 1960 and 2010, with the area under each portion of

Table 1
Correlation table: human development indicators.

	Life expectancy	IMR (ln)	Public health index
Life expectancy		−0.9025	0.9753
IMR (ln)	−0.9025		−0.9753
HALEs (males)	0.8157	−0.7810	0.8179
HALEs (females)	0.8225	−0.7972	0.8302
DALYs	−0.9416	0.7761	−0.8800
Child mortality rate (ln)	−0.9263	0.9958	−0.9853
Malnutrition (height for age)	−0.6630	0.7790	−0.7394
Poverty headcount ratio ($2/day)	−0.7560	0.8021	−0.8145
Human development index (HDI)	0.9181	−0.9320	0.9513

Pearson's r correlations, based on varying samples (no imputed data). All are significant at 99%. See Table A1 for variable definitions and sources.

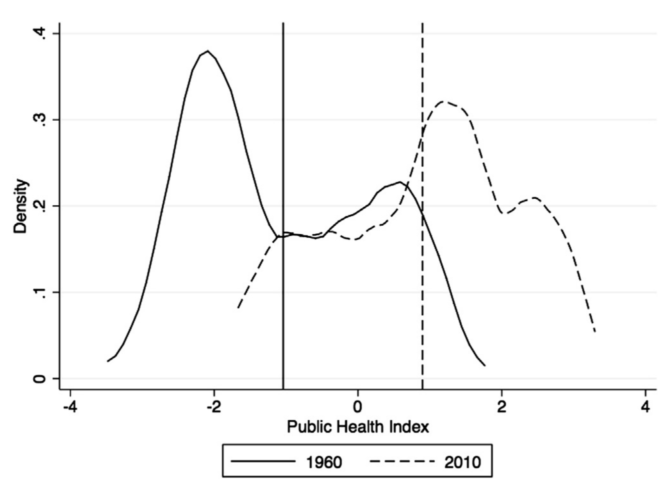

Fig. 1. Global distribution of public health by country. Kernel density plot of the distribution of public health in 1960 and 2010. Vertical line: mean value for that year's distribution. Unit of analysis: countries.

the curves indicating the share of countries with a corresponding level of health. The standard deviation of these distributions decreases only slightly from 1960 to 2010. However, the arithmetic mean increases dramatically from 1960 to 2010 (from −1.04 to 0.89), as indicated by the vertical lines in Fig. 1. By this measure, immense progress in human wellbeing has been accomplished over the past five decades (Gerring, 2007b; Goesling & Firebaugh, 2004; Neumayer, 2003).

This general conclusion is validated by alternative approaches to the same data. Note that because the population of sovereign countries changes considerably over the observed time-period, with many new countries coming into formal existence, the sample in 1960 ($N = 111$) is considerably smaller than the sample in 2010 ($N = 190$). However, the changing composition of the sample has only minimal impact on the shape of the density functions.

One might also display the distribution of public health after weighting each country by population or by separating life expectancy from IMR. Each representation of the data shows a somewhat different picture of changes over this 5-decade period, as depicted in Figs A1–A3. Yet, dramatic improvements in human welfare from 1960 to 2010 are registered for each indicator.

Despite these improvements, enormous differences continue to characterize the global distribution of health, as captured by the tails of the distribution in Fig. 1. One's chances of survival depend largely upon where one happens to be born. In Sierra Leone, in 2010, 113 out of every 1000 babies died before they reached their first year of life, and life expectancy was only 47 years—levels that approximate mortality rates among pre-modern populations. In Singapore, by contrast, the infant mortality rate was about 2.1 per 1,000, and life expectancy was 81 years. How are we to understand these differences?

A model-based approach to measuring health sector performance

Having constructed an index of public health we turn to the question of health sector performance. The health *sector*, or system, includes "all the activities whose primary purpose is to promote, restore, or maintain health" (WHO, 2000). To measure health sector performance, one must find a way to measure a country's public health against the background of its endowments—those factors affecting public health that the health sector cannot reasonably be held responsible for.

To conduct this discounting exercise we employ an ordinary least squares regression model,

$$Y = a_1 + b_1X_1 \ldots b_kX_k + \varepsilon \quad [1]$$

where Y = public health index and X_{1-k} = variables measuring various endowments, along with a vector of year fixed-effects. The latter are included in order to account for variation through time in global factors (commodity prices, global economic performance, the diffusion of ideas and technology) that affect all countries more or less equally.

Regression analysis is employed here for descriptive purposes rather than for causal inference. Accordingly, the coefficients and standard errors for each independent variable are of peripheral interest. We stake no claims about the validity or precision of these estimates. Our claim is simply that the model *as a whole* captures most of each country's endowments. Theoretical interest is focused on the residual from this model—the unexplained variance (ε)—which serves as a proxy for health performance in a country. The logic of this method of indirect measurement is illustrated in Fig. 2, where endowments likely to affect public health (but not the

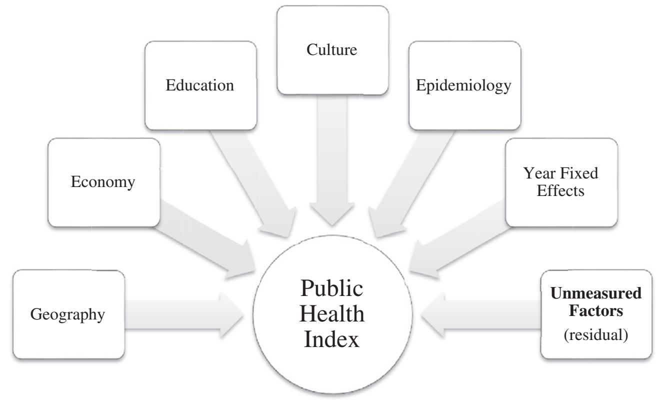

Fig. 2. A residual model of public health performance.

responsibility of the health sector) are divided into five areas—*economics*, *education*, *culture*, *geography*, and *epidemiology*.

Economics is captured by urbanization and per capita GDP. Our expectation is that richer countries and those with higher levels of urbanization should experience better health outcomes, all else being equal. We also include two polynomial terms for income (GDPpc squared and GDPpc cubed) so as to capture non-linearities in the relationship between economics and health.

Education is measured by the adult literacy rate (% of people above 15) and by school attainment (total years of schooling for people 25 and older).

Culture is operationalized, first, by variables measuring the share of a country's inhabitants whose religious background is (a) Catholic, (b) Protestant, or (c) Muslim. Second, we measure each country's legal culture with dummies for (a) English legal origin, (b) French legal origin, (c) German legal origin, and (d) Scandinavian legal origin.

Geography includes a variety of physical factors that might impact the spread of disease and other aspects of the physical quality of life: distance from the equator (transformed by the natural logarithm), island (dummy), landlock (dummy), and a series of climatic zone variables measuring the percent of a country's territory classified as polar, boreal, temperate desert, sub-tropical, tropical, wet temperate, tropical desert, and water. We expect that countries at a greater remove from the equator, situated on islands, with ocean access, and with more favorable climactic conditions will experience better health outcomes.

Epidemiology refers to the international vector of an epidemic disease, i.e., that portion of an epidemic that a country is exposed to by virtue of diffusion (and that it can do relatively little to control). This is represented in our model by the prevalence of (a) HIV/AIDS, (b) malaria, and (c) tuberculosis in neighboring countries. Specifically, the disease exposure for *Country A* is calculated as the mean value of disease prevalence in all countries (not including *A*): (1) whose capitals lie within 1600 km of *A*, or (2) whose borders are contiguous or nearly so (e.g., bodies of land separated by small bodies of water). The intuition is that if a country is surrounded by other countries with high rates of infection its exposure rate is correspondingly high. Accordingly, if a country maintains a lower (higher) rate of infection than its neighbors, it is judged a success (failure) in our model-based assessment.

To reiterate, the goal of this model-based exercise is not to provide a complete causal account of health outcomes. We purposefully exclude potential causal factors that lie within the health sector and are within a government's immediate control (i.e., that are endogenous to politics and policy making). Any attempt to measure such factors directly would result in a partial and probably biased account of public health performance, for reasons already noted. Instead, we focus on the unexplained variance captured in the model's residuals.

Prior to analyzing the data we correct for possible biases caused by missing data, which may produce a truncated, unrepresentative sample. Imputation of missing observations is carried out with the "Amelia II" algorithm (Honaker, King, & Blackwell, 2012) using data from all variables included in the following analyses and specifications that capture trends over time and within cross-sectional units. Five sets of estimates of the missing data are created. Variation across the imputed values in each dataset is then used to take into account the relative precision of the imputation procedure during subsequent analyses. Subsequent analyses are conducted with these imputed datasets, which include all sovereign and semisovereign countries from 1960 to 2010 ($N = 8099$). Descriptive statistics for all variables are shown in Table A2.

The chosen estimator for our model is ordinary least squares. OLS is preferred to models that allow flexible parameters such as hierarchical models with varying coefficients (e.g., mixture models) because of the special purpose of our model. Recall that we employ this model to generate an estimate of public health performance given certain background conditions. If we allow these background conditions to assume a country-specific shape, its shape will depend upon matters like health care. In this fashion, a model with flexible parameters will soak up variation that we wish to measure in the residuals.

Table 2 shows results from the model summarized in Equation (1), in which the public health index is regressed against endowments along with year fixed-effects. The fit is quite good ($R^2 = 0.826$), suggesting that baseline factors account for a substantial amount of the variance in public health. We refer to this as a *cross-sectional* model because it indicates the public health achievement of countries relative to each other at particular points in time (note the inclusion of year fixed-effects).

Table 2
A cross-sectional model of public health performance.

HIV exposure	−0.0352	[0.002]	***
TB exposure	−0.0003	[0.000]	***
Malaria exposure	−0.0000	[0.000]	***
GDP per capita	−1.7235	[0.247]	***
GDP per capita squared	0.2492	[0.032]	***
GDP per capita cubed	−0.0097	[0.001]	***
Urban population	0.0049	[0.000]	***
School attainment	0.0584	[0.009]	***
Adult literacy rate	0.0103	[0.002]	***
Catholic	0.0001	[0.000]	
Muslim	−0.0034	[0.000]	***
Protestant	−0.0027	[0.000]	***
English legal origin	−0.0157	[0.028]	
French legal origin	−0.1022	[0.025]	***
German legal origin	0.0684	[0.046]	
Scandinavian legal origin	0.4185	[0.058]	***
Island	0.2155	[0.019]	***
Distance from equator (ln)	0.0786	[0.014]	***
Landlocked	−0.0123	[0.020]	
Boreal	0.0061	[0.100]	
Temperate desert	−0.6274	[0.099]	***
Sub-tropical	−0.2794	[0.055]	***
Tropical	−0.0582	[0.050]	
Wet temperate	0.2078	[0.050]	***
Tropical desert	0.1444	[0.055]	***
Water	0.4118	[0.142]	***
Constant	1.5031	[0.584]	**
Year fixed effects	Yes		
Countries	190		
Years	1960–2010		
Observations	8099		
R^2	0.8257		

Cross-sectional model of public health performance, in which the public health index is regressed against all factors unrelated to the health sector. The excluded climate variable is Polar. All variables defined in Table A1. Standard errors in brackets. *sig. at 10%; **sig. at 5%; ***sig. at 1%.

Among the various factors represented in the model, economic variables loom large. Of course, we do not mean to suggest that development *by itself* explains public health. The economic terms in our equation likely represent much more than income and demography. We presume that they are also playing a proxy role for various correlates of modernization that have a direct or indirect impact on public health. To the extent that these factors co-vary empirically with economic development they are correctly understood as integral to the secular-historical process of modernization. For example, if countries tend to adopt more extensive and effective social policies as they develop economically, this fact should be reflected in our baseline model. By the same token, any deviations from the norm—perhaps by virtue of spending more or less than they "should" (given their level of development), or by virtue of allocating money more or less efficiently to public health problems—will be reflected in the residual for this model. This is what we mean by gaging country performance in the health sector relative to endowments.

A second econometric issue is the possible endogeneity that may exist between GDP per capita and public health. A country's health performance may influence its rate of economic growth—and therefore per capita GDP—by affecting productivity, labor supply, and the accumulation of human capital. The magnitude and direction of the effect, as well as the causal mechanisms linking the two, are debated in the literature, but insofar as health influences prosperity the coefficient for GDP per capita may be biased. Yet, there are good reasons to suppose that this bias is relatively small (Weil, 2007) and—more importantly—that it is more or less equal across countries. If the bias is constant, the residuals from this analysis are still good indicators of the *relative* success of countries around the world in addressing public health.

Our purpose, in any case, is not to test the relationship between right- and left-hand variables. Thus, whether the coefficients reported in Table 2 are precisely estimated is not of great concern; we assume that they are only approximations of some underlying data-generating process. Since the regression model is being used as a measurement technique rather than a technique of causal inference many of the identification issues that plague regression-based analysis are only minimally relevant here.

Our theoretical interest focuses on what the baseline model does *not* explain. Recall that we regard the residuals, the difference between the actual and predicted values of our public health index, as a measure of country performance in the health sector. That is, given a country's endowments at a particular point in time as well as global trends (captured by year fixed effects), how impressive is a country's health performance?

Over- and under-performers

Over- and under-performers in 2010, as identified by the cross-sectional model, are displayed in the left half of Table 3, along with their estimated residuals and 90% confidence intervals for each residual. This is referred to as Achievement because it indicates relative performance at a particular point in time.

The right half of Table 3 indicates countries that have improved the most, or the least, over the 1960–2010 period (or whatever period the country has been in existence). This is calculated by subtracting a country's residual in 2010 from its residual in the first year of the panel (1960 or its first subsequent year of formal sovereignty). This is labeled Improvement, as it registers change through time in a country's residual.

Table 3
Over- and under-performers.

Achievement (2010)			Improvement (1960–2010)		
Best performers					
1.	Syria	1.17 ± 0.09	1.	Maldives	1.70 ± 0.21
2.	Vietnam	0.93 ± 0.10	2.	Oman	1.07 ± 0.19
3.	Bangladesh	0.88 ± 0.11	3.	Nepal	1.06 ± 0.22
5.	Nepal	0.86 ± 0.12	4.	Nicaragua	1.05 ± 0.17
6.	Malaysia	0.82 ± 0.11	5.	Vietnam	1.01 ± 0.22
7.	Singapore	0.76 ± 0.11	6.	United Arab Emirates	1.01 ± 0.20
8.	Portugal	0.73 ± 0.11	7.	Saudi Arabia	0.98 ± 0.20
9.	Morocco	0.70 ± 0.11	8.	Malawi	0.92 ± 0.22
10.	Mozambique	0.69 ± 0.14	9.	Madagascar	0.92 ± 0.17
11.	Nicaragua	0.67 ± 0.09	10.	Bangladesh	0.90 ± 0.20
12.	Thailand	0.65 ± 0.12	11.	Mozambique	0.90 ± 0.22
13.	Bosnia/Herzegovina	0.64 ± 0.11	12.	Tunisia	0.89 ± 0.18
14.	Honduras	0.62 ± 0.08	13.	Peru	0.85 ± 0.20
15.	Costa Rica	0.62 ± 0.09	14.	Honduras	0.85 ± 0.16
16.	Solomon Islands	0.61 ± 0.11	15.	Chile	0.84 ± 0.22
17.	Maldives	0.60 ± 0.11	16.	Syria	0.80 ± 0.19
18.	Iceland	0.60 ± 0.10	17.	Turkey	0.78 ± 0.21
19.	Egypt	0.60 ± 0.08	18.	Liberia	0.73 ± 0.19
20.	Malawi	0.59 ± 0.11	19.	Guatemala	0.70 ± 0.17
22.	Yemen, Rep.	0.58 ± 0.13	20.	Egypt	0.70 ± 0.18
Worst performers					
164.	Haiti	−0.45 ± 0.10	170.	Philippines	−0.62 ± 0.18
167.	United States	−0.49 ± 0.11	171.	St. Vincent & Grenadines	−0.63 ± 0.23
168.	Azerbaijan	−0.50 ± 0.12	172.	Romania	−0.64 ± 0.19
169.	Lebanon	−0.50 ± 0.10	173.	Panama	−0.66 ± 0.18
170.	Jamaica	−0.53 ± 0.13	174.	Nigeria	−0.68 ± 0.24
172.	Congo, Dem. Rep.	−0.57 ± 0.10	175.	Jamaica	−0.69 ± 0.23
173.	Bolivia	−0.57 ± 0.12	176.	Uganda	−0.75 ± 0.19
174.	Kazakhstan	−0.61 ± 0.14	177.	Burundi	−0.75 ± 0.19
175.	Botswana	−0.67 ± 0.15	178.	Equatorial Guinea	−0.77 ± 0.19
177.	Cameroon	−0.77 ± 0.09	179.	Netherlands	−0.79 ± 0.19
178.	The Bahamas	−0.77 ± 0.10	180.	Lebanon	−0.79 ± 0.18
179.	Gabon	−0.77 ± 0.09	181.	Cameroon	−0.79 ± 0.18
182.	Nigeria	−0.86 ± 0.13	182.	Paraguay	−0.80 ± 0.19
183.	Angola	−0.88 ± 0.12	183.	Kenya	−0.81 ± 0.18
185.	Sierra Leone	−0.90 ± 0.13	185.	Bulgaria	−0.86 ± 0.20
186.	Trinidad/Tobago	−0.92 ± 0.11	186.	Zimbabwe	−0.91 ± 0.21
187.	South Africa	−1.00 ± 0.13	187.	Congo, Rep.	−0.95 ± 0.19
188.	Djibouti	−1.11 ± 0.10	188.	Trinidad and Tobago	−1.01 ± 0.20
189.	Swaziland	−1.17 ± 0.13	189.	Botswana	−1.09 ± 0.24
190.	Equatorial Guinea	−1.80 ± 0.11	190.	Russia	−1.12 ± 0.21

Achievement: rank, country name, residual, and 90% confidence interval for selected countries in 2010. *Improvement*: rank, country name, Δresidual from 1960 (or independence) to 2010. Results based on cross-sectional model of public health performance (Equation (1), Table 2). Countries listed subject to exclusions, as explained in the text.

While these residuals, like any others, capture measurement and stochastic error, we do not anticipate any systematic sources of error. A high positive residual indicates that a country over-performs in that year while a large negative residual indicates underperformance, relative to the parameters of the model. Note that we are less interested in the absolute size of the residuals (which are of course affected by any biases in the model) than in the placement of country residuals *relative to each other*.

Although models in Table 2 are derived from the entire population of sovereign and semisovereign countries (including some imputed data), countries listed in Table 3 are restricted to a subset that meet more demanding data requirements. First, at least 30 years of real (non-imputed) data must exist for our public health index and these data must include the final year in our analysis (2010). Second, reliable GDP per capita data must be available for the most recent decade. These two criteria reduce the potential sample of countries from 190 (the total number of cases in our imputed sample) to 172. While recognizing the utility of having the largest (and most representative) dataset for drawing inferences about the world, we do not want to make point predictions for specific countries using very poor data. This would be irresponsible and potentially misleading, especially since some users may disregard the unusually high confidence intervals for those countries.

Many of the over-performers listed in Table 3 are familiar to policy experts. Other countries—such as Egypt, Guatemala, Honduras, Liberia, Madagascar, Malawi, Maldives, Morocco, Mozambique, Nicaragua, Peru, Oman, Solomon Islands, Syria, and Yemen—are more surprising, suggesting that further research on these cases may be warranted. Most of the under-performers require little comment, and almost all are located in the developing world. Only one OECD country—the United States—qualifies as an under-achiever and an under-improver.

All historical-cultural regions of the world are represented among over- and under-performers. Indeed, region is a surprisingly weak predictor of country performance—with two notable exceptions. Countries in sub-Saharan Africa are more likely to be found among under-performers and countries in the Middle East are more likely to be found among over-performers. While the former story is well known, the latter story is largely unsung and surely deserves more attention.

While differences across regions naturally attract attention, for analytic purposes one is well-advised to focus on varying performance *within* a given region. Comparisons of this sort embody a "most-similar" case design (Gerring, 2007a).

It will be apparent that the lists of Achievers and Improvers in Table 3 overlap but are by no means identical. Achievers and Improvers in 2010 correlate at 0.5893 (Spearman's r) and 0.6063 (Pearson's r) across the full sample. Note that countries like Singapore, a high Achiever in 2010 but a middling Improver, made their biggest gains in public health prior to 1960, and thus do not register substantial improvement over the observed time-period. Evidently, one's choice of beginning and end-points determines results produced by any measure of improvement.

One might wonder how stable Achievement scores are through time. For this purpose it is informative to compare residuals from the Achievement model in 2010 (displayed for selected countries on the left side of Table 3) and in 1960 (or the year in which a country becomes independent)—the first and last years in our panel. These two sets of residuals are correlated at 0.3070 (Spearman's r) and 0.3646 (Pearson's r). Although statistically significant, this is a remarkably low correlation, suggesting that global leadership in public health performance has changed dramatically over the past half-century.

How much difference do our model-based adjustments make in judgments of country performance relative to "raw" (unadjusted) indicators? This may be assessed by comparing Achievement and Improvement residuals with the *unadjusted* public health index. Achievement residuals are correlated with the public health index at 0.2918 (Spearman's r) and 0.3177 (Pearson's r). Improvement residuals from 1960 to 2010 are correlated with changes in the public health index over the same period at 0.8357 (Spearman's r) and 0.8273 (Pearson's r). Clearly, a model-based approach to health differs substantially from a simple measure of the outcome of interest, though the differences are greater when looking at a single point in time then when looking at change over time, as one might expect.

Tracking within-country change: a country fixed-effect model

In some circumstances policymakers and publics are concerned primarily with a country's progress (or regress) over time, rather than its performance relative to other countries. Where temporal comparisons are more important than spatial comparisons we propose a modified version of Equation (1) in which Y = public health index and X_{1-k} = time-varying variables measuring various endowments, along with year and country fixed effects. Note that in this model dummy variables for each country (minus one) are inserted into the benchmark model and time-invariant variables (e.g., geographic factors) are excluded. This controls for characteristics of a country that do not vary over time and simplifies the specification problem. We refer to this as a *country fixed-effect* approach to model specification. Results are displayed in Table 4.

Residuals generated in this fashion are most useful when examined over the entire observed period. Note that country performance in public health, as with many policy areas, is rarely uniform over time. In order to get a sense of the temporal variation in performance realized by a single country over this five-decade period we provide a graph of residuals for Egypt, one of the most-improved countries in our sample. Of particular interest in Fig. 3 are the changes in trend from decade to decade and the dramatic improvement in health performance (relative to background factors) that began in the 1980s. While different functional forms and estimation techniques might generate slightly different tendencies, the overall longitudinal pattern found in graphs of this nature can be extraordinarily informative insofar as they allow for a sensitive tracking of country performance over time, while holding constant background factors that lie outside the health sector.

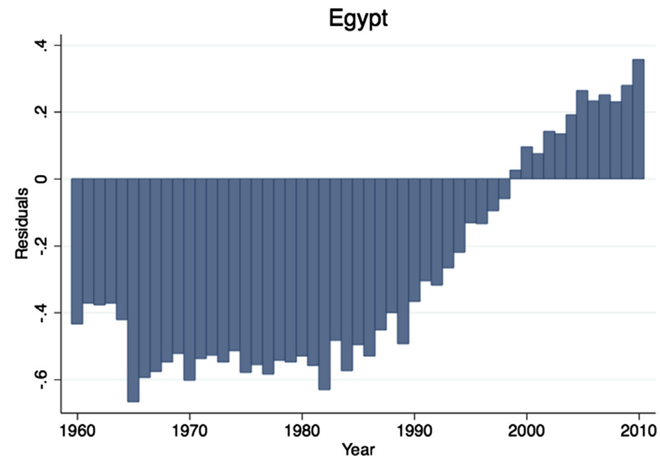

Fig. 3. Residuals from the country fixed-effect model: an example.

Robustness tests

To reiterate, the findings of interest in this study are contained in the residuals for each country, not in the coefficients and standard errors attached to specific parameters. Of course, we must be concerned if a variable's performance confounds theoretical expectations. However, we are not concerned to arrive at precise parameter estimates. The purpose of this statistical exercise is therefore quite different from the usual employment of statistical models, which is to test a general causal relationship.

We are still concerned with robustness. However, our concern centers on the behavior of the residuals—which we regard as a measure of health sector performance—rather than the parameter estimates. Would Syria still be the highest achiever, as shown in Table 3, if aspects of the model were altered? Would the Maldives still be the greatest improver? To what extent would these residuals, and the relative ranking of countries, be affected by minor changes in modeling strategy?

To test the stability of our findings we run a series of sixteen robustness checks for the two benchmark models—the cross-sectional model (Table 2) and the country fixed-effect model (Table 4). Each robustness test involves a single change in a benchmark model. We then compare the resulting residuals (between the benchmark model and the slightly altered model) according to two metrics: the rank correlation (Spearman's r) of the residuals and the value correlation (Pearson's r) of the residuals. Results are displayed in Table 5.

First, we test how the construction of the public health index affects the results. In row (1), we examine the correlation between residuals produced by the benchmark models and two additional

Table 4
Country fixed-effect model of public health performance.

HIV exposure	−0.0363	[0.002]	***
TB exposure	−0.0003	[0.000]	***
Malaria exposure	−0.0000	[0.000]	
GDP per capita	−0.6798	[0.233]	***
GDP per capita squared	0.1261	[0.031]	***
GDP per capita cubed	−0.0054	[0.001]	***
Urban population	0.0058	[0.001]	***
School attainment	0.0145	[0.005]	***
Adult literacy rate	0.0033	[0.001]	***
Constant	−1.0485	[0.563]	*
Year fixed effects	Yes		
Country fixed effects	Yes		
Years	1960–2010		
Countries	190		
Observations	8099		
R^2 *(within)*	0.8063		

Country fixed-effect model of public health performance, in which the public health index is regressed against time-varying factors along with country and year fixed effects. All variables defined in Table A1. Standard errors in brackets. *sig. at 10%; **sig. at 5%; ***sig. at 1%.

Table 5
Robustness tests.

	Cross-sectional model		Fixed-effect model	
	Spearman's r	Pearson's r	Spearman's r	Pearson's r
I. Alternative outcome measures				
1. Life expectancy only	0.8821	0.8963	0.9347	0.9388
2. IMR (ln) only	−0.8501	−0.8566	−0.8824	−0.8605
II. Subtracting regressors				
3. −HIV, TB, malaria prevalence	0.9296	0.9326	0.9762	0.9706
4. −ln(GDP) & its polynomials	0.8743	0.8776	0.9506	0.9482
5. −Urban population	0.9879	0.9894	0.9880	0.9886
6. −Geographic controls	0.9428	0.9490	–	–
III. Adding regressors				
7. +Growth (GDP per cap)	1.0000	1.0000	1.0000	1.0000
8. +Health expenditure per cap	1.0000	1.0000	0.9963	0.9964
9. +Imports (share of GDP)	0.9993	0.9994	0.9995	0.9996
10. +Oil production per cap	0.9919	0.9938	1.0000	1.0000
11. +Democracy stock	0.9906	0.9925	0.9995	0.9995
12. +Tax revenue	0.9765	0.9814	0.9990	0.9991
13. +Telephone mainlines	0.9710	0.9767	0.9999	0.9999
14. +Gini index	0.9936	0.9948	1.0000	1.0000
15. +Conflicts	0.9999	0.9999	0.9992	0.9993
IV. Changes to the sample				
16. −OECD countries	0.9444	0.9306	0.9978	0.9979

N = 8099 except row 16, where N = 6635.

models that are identical in all respects except that life expectancy (rather than the public health index) forms the dependent variable. In row (2), we repeat this exercise by comparing the benchmark models with those in which IMR alone forms the dependent variable.

In the second set of tests (rows 3–6), we explore the effect of removing various factors from the right side of the benchmark models: (3) HIV, TB, and malaria prevalence, (4) GDP & its polynomials, (5) urban population, and (6) the geographic controls. After removing each set of variables (seriatim), we correlate the residuals from that model with the results of the benchmark model.

The third set of models (rows 7–15) adds other theoretically plausible variables into the benchmark models: (7) GDP per capita growth, (8) health expenditure per capita, (9) imports (as a share of GDP), (10) oil production per capita, (11) democracy stock (a measure of a country's regime history going back to 1900), (12) tax revenue (as a share of GDP), (13) telephone mainlines, (14) Gini index of income inequality, and (15) conflicts (intra- and inter-country). Again, the residuals produced by these re-specified models are compared with the residuals from the benchmark models.

Finally (row 16), we adjust the sample, excluding OECD countries. This tests the possibility that fundamentally different causal relationships may exist in the developing and developed worlds.

Table 5 indicates that the results of our benchmark models are somewhat affected by an alteration in the outcome (rows 1–2) and by the subtraction of GDP and its polynomials (row 4). However, residuals produced by these robustness tests are still highly correlated with residuals produced by our benchmark models (all correlations are above 0.85). Other changes have even less impact (all correlations are above 0.90).

Even if observers cannot agree on precisely which elements belong in a benchmark model of public health performance, differences across alternate models are fairly small. This reinforces our sense that a model-based assessment of health performance is viable.

It is noteworthy that the inclusion of a growth term has virtually no impact on the relative placement of countries in the cross-sectional and fixed-effect models (row 7). This reiterates the message contained in Table 2, where we find that among over-performers are countries with stagnant economies (e.g., Egypt, Syria, Tunisia) as well as dynamic economies (e.g., Chile, Saudi Arabia, Singapore, UAE, Turkey, Vietnam) in the postwar era. There appears to be no consistent tradeoff between growth and health.

The point is vital for the underpinnings of our model. Consider that if achieving growth and health were simply a matter of allocating necessary resources one would expect different societies to make different choices. Some would be "capitalist" states, in which growth is prioritized over other goods; others would be "socialist" states, in which social development is granted priority. There is little empirical support for this zero-sum view of the policy world. This non-relationship is comprehensible when one recalls that there is no correlation between aggregate taxing and spending or social policy spending and growth performance. Big spenders, and big welfare states, do not appear to grow more or less slowly (Lindert, 2004). And this, in turn, reinforces our sense that social policy and economic policy are non-rival. Both can be pursued successfully at the same time. Alternatively, both can be simultaneously neglected, as the experience of many countries amply demonstrates.

Conclusions

An important step towards an effective solution to the myriad problems of international health is a meaningful metric of performance. In this paper we have suggested a methodology for measuring success and failure across countries with vastly different endowments, which we understand as encompassing economic, educational, cultural, geographic, and epidemiological components.

The proposed approach to health performance differs from extant approaches, which generally do not take into account varying endowments. To be sure, writers often employ a back-of-the-envelope procedure, which usually amounts to a comparison between one country's performance on a measure of human development and other, similarly situated, countries' performance on that same indicator. Costa Rica may be compared with other developing countries and Sweden may be compared with other advanced industrial countries. Our approach may be understood as an effort to systematize the intuitions behind these simple comparisons.

In concluding, we might recall that the core motivation of a model-based approach to public health performance is to provide a basis for judging success and failure by reference to things that citizens and policymakers could affect without any change in a country's background conditions. We do *not* purport to have provided a full causal model of health, which would necessarily include many additional variables not found in our benchmark models. Our approach is thus properly classified as a descriptive model with strong prescriptive overtones, and in this respect mirrors the goals of the Human Development Index (HDI), the Physical Quality of Life Index (Morris 1979), and various measures of policy efficiency (Evans, Tandon, Murray, & Lauer, 2001; Gupta & Verhoeven 2001; Moore 2003; reviewed in Ravallion 2003).

A residual-based approach to health does not, of course, shed light on the *reasons* for the unexplained variance, i.e., the reasons why some countries have positive residuals and others have negative residuals. We suspect that the relative efficiency and effectiveness of health systems in countries around the world play a large role in this story. Even so, the interpretation of a country's residual in our models is a complex matter. In claiming that a country over- or under-performs we do not intend to point the finger at any one source. The reason for over- or under-performance can only be understood through further analysis, presumably including in-depth country studies. Our hope is that the benchmark models provided in this analysis will provide a suitable tool for case-selection where scholars wish to conduct case studies, as well as an impetus for new theorizing, data gathering and empirical research on the causal determinants of public health performance. To take the lessons of this benchmarking exercise to the next logical step we will need a better understanding of the complex web of political, social and economic factors that, together, influence health outcomes. This includes the role played by political regimes, constitutional structures, the capacity and resources of the health sector, political corruption, economic policies, bureaucratic efficiency, and the role played by non-governmental and international organizations.

Our approach has important policy implications as well. Consider that countries with high residuals may provide exemplars—"best practices" that could be adapted for use in low-performing countries. In this respect, a model-based procedure may help to shed light on concrete options for reform.

Likewise, a country's residual may indicate its likely potential for improving public health and the possible payoff to be derived from further investments in the health sector. Although countries with high positive residuals have demonstrated a capacity for effective health policy and improvement, some high-performing countries may be bumping up against what it is possible to achieve, given current constraints (as measured by each country's endowments). By the same token, countries with negative residuals in the cross-sectional model could probably do a lot more with the resources currently available to them. They are not economically, educationally, culturally, geographically, or epidemiologically constrained, so far as we can tell. It is our hope that the use of model-based analyses of the health sector will assist in identifying targets of opportunity, informing national and international health agencies and advocates, and applying political pressure that might lead, ultimately, to improved performance.

Acknowledgment

We are grateful for comments and suggestions received from Kirk Deardon, Varun Gauri, Kenneth Hill, Dean Jamison, Gerald Keusch, Christopher Murray, Kandala Ngianga-Bakwin, Lant Pritchett, Michael Reich, Jon Simon, and participants at talks given at the Population and Development Center, Harvard University, and the Frederick S. Pardee Center for the Study of the Longer-Range Future, Boston University. This research was supported by the Provost's Office at Boston University and the Development Research Group and the Research Support Budget of the World Bank.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.socscimed.2013.06.002.

References

Backman, G., Hunt, P., Khosla, R., Jaramillo-Strouss, C., Fikre, B. M., Rumble, C., et al. (2008). Health systems and the right to health: an assessment of 194 countries. *The Lancet, 372*(9655), 2047–2085.

Caldwell, J. C. (1986). Routes to low mortality in poor countries. *Population and Development Review, 12*(2), 171–220.

Evans, D. B., Tandon, A., Murray, C. J. L., & Lauer, J. A. (2001). *Comparative efficiency of national health systems in producing health: An analysis of 191 countries*. (GPE Discussion Paper Series No. 29). Retrieved from World Health Organization website http://www.who.int/healthinfo/paper29.pdf.

Gerring, J. (2007a). *Case study research: Principles and practices*. Cambridge: Cambridge University Press.

Gerring, J. (2007b). Global justice as an empirical question. *PS: Political Science and Politics, 40*(1), 67–78.

Ghai, D. (Ed.). (2000). *Social development and public policy: Some lessons from successful experiences*. London: St. Martin's.

Goesling, B., & Firebaugh, G. (2004). The trend in international health inequality. *Population and Development Review, 30*(1), 131–146.

Gupta, S., & Verhoeven, M. (2001). The efficiency of government expenditure: experiences from Africa. *Journal of Policy Modeling, 23*(4), 433–467.

Halstead, S. B., Walsh, J. A., & Warren, K. S. (Eds.). (1985). *Good health at low cost: A conference report*. New York: The Rockefeller Foundation.

Honaker, J., King, G., & Blackwell, M. (2012). *AMELIA II: A program for missing data*. Cambridge, MA: Harvard University. Retrieved from http://GKing.Harvard.edu/.

Jamison, D. T., & Sandbu, M. E. (2001). WHO ranking of health system performance. *Science, 293*(5535), 1595–1596.

Kakwani, N. (1993). Performance in living standards: an international comparison. *Journal of Development Economics, 41*(2), 307–336.

Kruk, M. E., & Freedman, L. P. (2008). Assessing health system performance in developing countries: a review of the literature. *Health Policy, 85*(3), 263–276.

Lindert, P. H. (2004). *Growing public. Social spending and economic growth since the eighteenth century*. Cambridge: Cambridge University Press.

McGuire, J. W. (2010). *Wealth, health, and democracy in East Asia and Latin America*. Cambridge: Cambridge University Press.

Mehrotra, S., & Jolly, R. (Eds.). (1997). *Development with a human face: Experiences in social achievement and economic growth*. Oxford: Clarendon.

Moore, M. (2003). Polity qualities: how governance affects poverty. In P. P. Houtzager, & M. Moore (Eds.), *Changing paths: International development and the politics of inclusion* (pp. 167–203). Ann Arbor: University of Michigan Press.

Morris, D. (1979). *Measuring the condition of the world's poor: The physical quality of life index*. New York: Pergamon.

Murray, C. J. L., & Evans, D. B. (Eds.). (2003). *Health systems performance assessment: Debates, methods and empiricism*. Geneva: World Health Organization.

Ndulu, B. J., & O'Connell, S. A. (2007). Policy plus: African growth performance 1960–2000. In B. J. Ndulu, et al. (Eds.), *The political economy of economic growth in Africa, 1960–2000: Vol. 1*. (pp. 3–75). Cambridge: Cambridge University Press.

Neumayer, E. (2003). Beyond income: convergence in living standards, big time. *Structural Change and Economic Dynamics, 14*(3), 275–296.

Ravallion, M. (2003). *On measuring aggregate 'social efficiency'*. (World Bank Policy Research Paper 3166). Retrieved from http://elibrary.worldbank.org/content/workingpaper/10.1596/1813-9450-3166.

Riley, J. C. (2007). *Low income, social growth, and good health: A history of twelve countries*. Berkeley: University of California Press.

Wang, J., Jamison, D. T., Bos, E., Preker, A., & Peabody, J. (1999). *Measuring country performance on health: Selected indicators for 115 countries*. Washington: World Bank.

Weil, D. N. (2007). Accounting for the effect of health on economic growth. *Quarterly Journal of Economics, 122*(3), 1265–1306.

WHO. (2000). *World health report 2000: Health systems improving performance*. Geneva, Switzerland: World Health Organization.