



POLICY FORUM: GLOBAL HEALTH

WHO Ranking of Health System Performance

Dean T. Jamison* and Martin E. Sandbu

For over half a century, the availability of economic performance indicators—such as GDP per capita and inflation rates—has made it possible to hold political leaders accountable for economic management. Equally important, these economic outcome measures (and the entire system of national income and product accounts) have allowed evidence to supplant ideology for judging the soundness of alternative macroeconomic policies.

Publication of robust, transparent, and valid indices of health system performance could, likewise, lead to greater political accountability and to evidence-based health policies.

dEbate!

Respond online
<http://www.sciencemag.org/cgi/content/summary/293/5535/1595>

To this end the World Health Organization (WHO), in its *World Health Report 2000* (WHR2000), published indices of

health system performance for its 191 member states (1, 2). WHO's farsighted leadership has initiated a process that will ultimately improve the evidence base for health policy. We argue, however, that WHO's current performance algorithm has critical shortcomings and that the challenge of constructing valid measures remains.

Performance Measures

In its *World Health Report 1999* (WHR1999), WHO published measures by which country performance could be ranked relative to what would be predicted by income level (3). Rankings of health system performance would add substantially to knowledge of overall country performance for evaluating and improving health policy. In its WHR2000, WHO seeks to disentangle system performance from other determinants of health outcomes. The resulting rankings correlate only slightly with the 1999 country performance rankings—raising the questions of what types of performance measures can

meaningfully be constructed, and how, from available data.

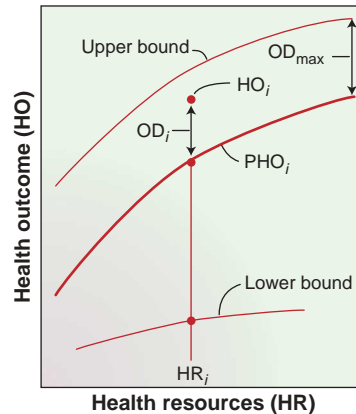
For many purposes, outcome levels will convey performance. That Canada's male life expectancy in 1999 of 76.2 years exceeded that of the United States by 2.4 years provides valuable information. When countries differ markedly in income, however, relative levels may prove more informative. The term "country performance" indicates how well a country is doing relative to what would be predicted from income and perhaps other specified determinants (4). Country performance indicators provide a starting point for discussion of why a country may be doing well or poorly.

That a country's performance is favorable could result from multiple factors: high levels of health expenditure, high efficiency in use of health expenditures (good health system performance), favorable geography, good governance, or luck. To assess health system performance, as opposed to country performance, requires identifying how outcome for each country responds to a change in inputs. WHR2000 simply assumes that system performance variation accounts for all outcome variation after controlling for levels of health expenditure and education. No outcome variation results from other determinants of health or from limitations in the underlying model.

As might be expected, the correlation between WHR2000 measures and country performance is low: Twenty out of 96 countries moved either up or down by 25 percentile points or more between the two rankings. For example, The Gambia's 1999 placement at the 9th percentile contrasts with its 2000 ranking at the 50th (5). Likewise the WHR2000 ranking correlates negatively with citizens' satisfaction with their own health care system for 17 high-income countries of the Organization for Economic Cooperation and Development (OECD).

Only 20% of Italians rate their health care system as satisfactory although Italy is number 2 in the WHR2000 ranking. Denmark, ranked 16 out of the 17 by WHR2000, had 91% of its citizens convey satisfaction (6).

Given the dramatic differences between rankings, it is important to examine how they are derived (7, 8). The figure on this page illustrates the WHR2000 methodology.



Measuring performance of health systems in WHR2000. The heavy line plots a statistical prediction of health outcome as a function of health resources. Abbreviations and the relation between the prediction line and the bounds are explained in the main text (13).

The heavy line in the middle traces out the health outcome that would be statistically predicted (PHO) for a country given its health expenditures. The actual health outcome for country i will typically deviate from the predicted outcome, and this is illustrated by the placement of HO_i (actual health outcome in country i). The vertical distance $HO_i - PHO_i$ is the outcome deviation for country i , OD_i .

In WHR2000, the authors define performance in terms of an upper bound that would be achieved by a maximally efficient health

system and a lower bound that is "the least that could be demanded..." [p. 41 in (1)]. "Maximal efficiency" was assumed to be the prediction line plus the maximum outcome deviation, for any country, OD_{max} . The lower bound somehow emerges from early 20th-century data on today's high-income countries.

Specifically, the WHR2000 index indicates how far a country is above the lower bound (LB_i) as a fraction of the distance between the upper and lower bounds:

$$\text{country performance} = [OD_i + (PHO_i - LB_i)] / [OD_{max} + (PHO_i - LB_i)]$$

In the two-dimensional context of the figure, the WHR2000 method can be viewed as a country-specific correction of the upper bound to adjust for education. As with country performance, health system performance becomes a function of OD_i , albeit a complex and nontransparent one.

Three points are clear: Including additional determinants will explain more of the outcome deviation, but by different amounts in different countries. Hence, the relative sizes of the residuals attributed to health system efficiency could be markedly affected. Second, even with all plausible controls included, shortcomings of model and data imply that remaining outcome de-

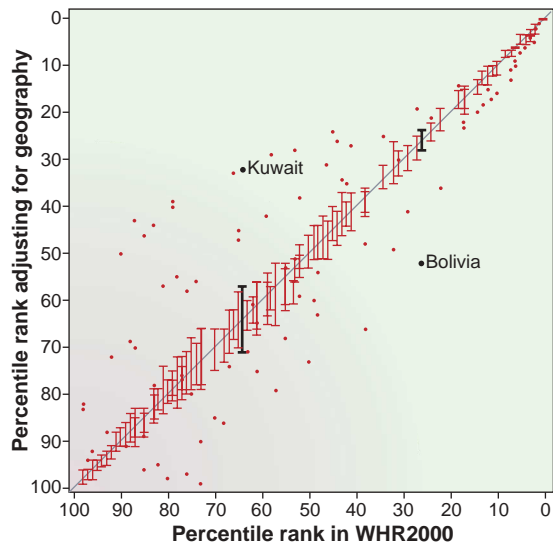
D. T. Jamison is with the Program on Global Health and Education, School of Public Health, University of California, Los Angeles, CA 90095, USA. M. E. Sandbu is in the Center for International Development, The Kennedy School of Government, Harvard University, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed.
 E-mail: djamison@isop.ucla.edu

viation will result only partly from efficiency variation. Third, multiple arbitrary assumptions define the upper and lower bounds, leaving rankings sensitive to these assumptions.

Sensitivity Analysis

How quantitatively important are these concerns? We assessed this with a sensitivity analysis by adjusting the upper bound for geography, i.e., we added geographical variables to health expenditures and education in predicting outcomes. (Tropical locations, for example, appear to affect health adversely.) We correspondingly adjusted the upper bound, then recalculated rankings (9, 10).



Effect of adding geographical variables to WHR2000 rankings (13). The dots relate a country's WHR2000 percentile rank on health outcomes to the rank generated from a sensitivity analysis that uses WHR2000's methods but controls for geographical features of the country. Uncertainty intervals (vertical bars) indicate WHR2000 authors' confidence in the ranking for each country. For example, Bolivia's range was from the 25th to 29th percentiles (black bar).

The figure on this page plots each country's geography-adjusted percentile rank against the WHR2000 rank. Rankings coincide along the 45° line. Vertical bars display the "uncertainty interval" for each country's ranking as given in WHR2000 (in percentiles). For only 17 out of 96 countries does the geography-adjusted rank fall within the uncertainty interval. Bolivia, for example, performed poorly according to WHR2000, ranking at the 26th percentile. After adjusting for geography, however, Bolivia's percentile increases to 52, i.e., by more than six times the width of its uncertainty interval (see figure, this page). This suggests that the uncertainty intervals convey no information or, worse, that they misleadingly convey precision (11).

We are not claiming that the WHR2000 algorithm would generate meaningful rank-

ings if geographical controls were added. This analysis only points to great sensitivity in results when variables (in addition to education) are controlled. The conceptual problem is more fundamental. Some of a country's outcome deviation results from how well its health system performs. It could, for a particular country, be 10% of the outcome deviation, it could be 50%, it could be anything. WHR2000 assumes 100% for all countries, because it lacks a way of estimating the actual values. We also examined the sensitivity of the rankings to different percentage assumptions, and as with the findings on geography, the rankings differ markedly. The authors of WHR2000 offer no empirical justification for their assumption of 100%.

Empirical Assessment of Performance

Is it possible to estimate the amount of a country's outcome deviation that is, indeed, attributable to how efficiently it uses resources? Time series data on health expenditures, in addition to the other relevant variables, would allow an attempt to estimate statistically the responsiveness (or elasticity) of health outcomes with respect to health expenditures separately for each country. If such an approach succeeded, it would provide one potential empirical approach to defining health system performance. Methods of multilevel modeling used in education research provide examples for proceeding (12).

Even the more modest objective of assessing country performance requires substantial caveats. That said, country performance measurement could be substantially improved over what WHR1999 reported. In particular, measures of country performance relative to geography, in addition to income and education, are probably close to the best that can be done without time series measures of health resources.

Conclusion

In the past several years, the World Bank and WHO have published quantitative measures of country performance in health. By highlighting its rankings in WHR2000, and by attributing the results to differences in the efficiency of health systems, WHO attracted extensive media attention. WHR2000 has both stimulated and contributed to a much-needed debate.

High visibility runs the risk, however,

of a counterproductive effect if technical mistakes remain uncorrected and resultant rankings unsupportable. Directions suggested in this note will, we hope, contribute to the evolution of WHO's important performance measurement initiative.

References and Notes

1. WHO, *World Health Report 2000—Health Systems: Improving Performance* (World Health Organization, Geneva, 2000).
2. WHR2000 provides rankings of health systems with respect both to health (disability-adjusted life expectancies or DALES) and overall performance. Overall performance combines measures of attainment not only on health but also on aspects of finance, responsiveness, and distribution. This policy forum discusses performance with respect to health, but the same issues arise concerning overall performance. For a critical discussion of the content and construction of all these measures see A. Williams, *Health Econ.* **10**, 93 (2001).
3. WHO, *World Health Report 1999: Making a Difference* (World Health Organization, Geneva, 1999). Its Annex table 6 reports country performance relative to income between 1952 and 1992 on infant mortality rate and female life expectancy. One author of this policy forum (D.T.J.) was the lead author of WHR1999.
4. Similar analysis is possible at the level of hospitals rather than countries. California, for example, assessed performance differences in management of acute myocardial infarction relative to variables that control for the complexity of the case mix presenting at the hospital. [See H. S. Luft, P. S. Romano, Principal Investigators, *Second Report of the California Hospital Outcomes Project*, vols. 1 and 2 (California Office of Statewide Health Planning and Development, 1996). Hospitals differed markedly in risk-adjusted performance.
5. The percentiles refer to the countries' WHR1999 and WHR2000 ranks, respectively, relative to the subset of countries included in both rankings. WHR1999 used level of female life expectancy in 1992 as its outcome measure, whereas WHR2000 used DALE in 1997. Data on life expectancy and DALE are virtually perfectly correlated, so use of DALES imposes the cost of lower transparency for virtually no gain, and in particular, this is not a source of difference between the WHR2000 and the WHR1999 rankings. (Female life expectancy in 1992 has a 0.95 correlation with DALES in 1997.)
6. See R. J. Blendon, M. Kim, J. M. Benson, *Health Affairs* **20**, 10 (2001).
7. The WHR2000 algorithm was proposed by C. J. L. Murray and J. Frenk, *Bull. W.H.O.* **78**(6), 717 (2000).
8. The method is described in D. Evans, A. Tandon, C. J. L. Murray, J. A. Lauer, "The comparative efficiency of national health systems in producing health: An analysis of 191 countries" (Global Program on Evidence discussion paper no. 29, World Health Organization, Geneva, 2000).
9. Because the actual frontier values for each country were not published by WHO, we approximated using information in D. B. Evans *et al.* (8). We regressed the measure of health outcome on the same variables they used plus two geographical variables provided by J. L. Gallup and J. Sachs, *Brookings Pap. Econ. Activity* **2**, 207 (1998).
10. D. Bloom and J. Sachs (9) find that tropical latitudes and geographical isolation adversely affect countries' growth prospects because of less-productive agriculture, a more hospitable environment for major diseases, and more difficulty in intellectual contact with the rest of the world.
11. WHR2000 includes exceptionally precise estimates of the ranks of Tanzania and Uganda: Both countries were reported to lie within an uncertainty interval including only ranks 179 and 180.
12. For a discussion of these methods, see A. S. Bryk and S. W. Raudenbush, *Hierarchical Linear Models* (Sage Publications, Newbury Park, CA, 1992).
13. The first and second figure are based on figures 2 and 9, respectively, of (8).
14. The authors appreciate valuable input from G. Alleyne, E. Bos, W. Hsiao, P. Jha, P. Musgrove, S. Nguyen, X. Pitkow, J. Sachs, and J. Wang.