

## Health metrics and evaluation: strengthening the science



Christopher J L Murray, Julio Frenk

With the growing importance of health in the global agenda comes the responsibility to develop a scientific foundation of metrics and evaluation. The scope of this emerging field can be viewed in terms of key topics, including health outcomes, other social outcomes related to health systems, health services, resource inputs, evaluations of programmes and systems, and analyses to support policy choice. It can also be defined in terms of key activities that are needed to strengthen the scientific basis of the field: development of new methods, instruments, software, and hardware; setting global norms and standards for data collection; increasing the availability of high-quality primary data; systematic analysis and synthesis of existing datasets; strengthening national capacity to obtain, analyse, and use data; and reporting and disseminating results. We explore in depth topics with major scientific challenges and institutional and cultural barriers that are slowing the development of the field. Cutting across the various topical areas and disciplinary approaches to these problems are some common scientific issues, including limited comparability of measurement, uncorrected known biases in data, no standard approach to missing data, unrealistic uncertainty estimates, and the use of disease models that have not been properly validated. Only through concerted action will it be possible to assure the production, reproduction, and use of knowledge that is crucial to the advancement of global health.

### Introduction

As the importance of health in the global agenda grows, so does the responsibility to measure accurately its complex dimensions and to assess the effects of increasing investments. The present burst of political and financial will to improve global health has to be matched by an adequate response from the community of experts to assure that the challenges are well understood and resources are applied in the best possible way. These goals can only be achieved if there is a firm foundation of metrics and evaluation. Indeed, the scientific approach to solving human challenges needs valid and reliable measurement combined with a systematic process to learn from experience through the accumulation of a body of knowledge.

Yet both clinical care and health-policy formulation too often do not have an adequate evidence base.<sup>1,2</sup> There is a paradox in this limitation, since the Latin roots of the words medicine (*mederi*, meaning to heal) and measurement (*metiri*) are the same<sup>3</sup>—namely, to take appropriate measures. In their origins, both terms were united by the common meaning of devising a rational course of action. We should recover this original unity to assure that taking measures to heal is always informed by taking the measurements that will guide action and assess results.

Efforts to introduce the discipline of measurement into the practice of medicine and public health are, of course, not new. For example, the work of the General Register Office in England beginning in the 1840s to publish comparative death rates by city, galvanised public-health action.<sup>4</sup> This demonstration of the power of measurement to catalyse reductions in child mortality came before the germ theory of disease. Recent years have seen a resurgence of interest in development of an evidence base for health policy,<sup>5-7</sup> which has to derive from rigorous measurement and evaluation. A clear example of such interest is the establishment of the Institute for Health

Metrics and Evaluation at the University of Washington, USA.<sup>8</sup> This and similar initiatives represent a constructive piece of the institutional architecture for global health, since they provide independent assessments at a time when the number of organisations, programmes, and partnerships is increasing.

More broadly, the field of metrics and evaluation can serve several purposes: first, to sustain interest in and funding for global health by demonstrating positive results; second, to enhance efficiency by building a solid knowledge base of what works, thus generating a process of shared learning between countries; third, to improve the quality of decision making by providing sound evidence; fourth, to foster interdisciplinary dialogue by bringing together various areas of enquiry; and last, to promote the values of transparency and accountability as essential ingredients of democratic governance both nationally and globally.

Health metrics and evaluation should not be seen in isolation of continuing work in related fields; it provides crucial inputs in the form of data and methods to areas, such as health-services research. For instance, comparative health-systems research needs valid measurements on performance, which can be generated by rigorous metrics. Similarly, research at the provider level can be enriched by methodological developments that are at the core of the area of health metrics and evaluation. The same relation applies to emerging fields, such as implementation science,<sup>9</sup> which is dependent on valid methods and data to measure success in translation of policy into action.

For this full potential to be realised, the scientific basis of health-policy formulation and programme implementation needs to be strengthened by the development of a rigorous basis of metrics and evaluation. We use the term metrics to refer both to the science of measurement and to the specific set of instruments and indicators that provide the empirical basis to understand

*Lancet* 2008; 371: 1191-99  
See [Comment](#) page 1139

Institute for Health Metrics and Evaluation, Seattle, Washington, USA (C J L Murray MD, J Frenk MD); Global Health, University of Washington School of Medicine, Seattle, Washington, USA (C J L Murray MD); Global Health Programme, Bill & Melinda Gates Foundation, Seattle, Washington, USA (J Frenk); and Carso Health Institute, Mexico City, Mexico (J Frenk)

Correspondence to: Christopher J L Murray, Institute Director, Institute for Health Metrics and Evaluation, Box 358210, 2301 5th Avenue Suite 600, Seattle, WA 98121, USA [cjlm@u.washington.edu](mailto:cjlm@u.washington.edu)

a particular object of enquiry and action (in our case, global health). In turn, the term evaluation refers to the disciplined attempt to establish a causal connection between an intervention and an effect. We see metrics and evaluation as an integrated scientific field that is interdisciplinary, bringing together notions, methods, and techniques from statistics, demography, epidemiology, economics, and other social sciences. Although most of the work outlined here builds on quantitative analyses of health policy, qualitative research has an important part to help understand contextual factors, to stimulate hypotheses, and to improve measurement instruments.

The purpose of this paper is to contribute to the scientific development of the field of metrics and evaluation. We map the range of topics and activities needed for the field and give special attention to six topics.

### What is the scope of required work?

The topics and types of activities that are needed to advance the science of metrics and evaluation can be characterised in many ways. One is to imagine a matrix with six broad topics and seven crucial activities that are required to lend support to analytical work. The broad headings for topics are: health outcomes, other social

outcomes related to health systems, health services, resource inputs, evaluations of policies and systems, and analyses to support policy choices. Within each of these, we have identified some subtopics that are not exhaustive but are meant to capture important components needed to track global-health conditions and the health system response (panel 1). We believe that inequalities are an important dimension of every topic, from child mortality to effective coverage of interventions, and to priority setting. We have also listed inequalities as a separate subtopic to draw attention to their importance.

For each topic, we identify seven activities that are integral to the advancement of global-health metrics and evaluation (panel 2). We believe that work for health outcomes is the most advanced, followed by that for programme evaluations and then financial resource inputs. Even in these arenas, however, further improvement is clearly needed. In terms of activities, we believe that work for setting global norms and standards other than the International Classification of Diseases (ICD) and the System of Health Accounts (SHA) is lagging behind, as is that for the availability of high-quality primary data and capacity strengthening.

In the matrix defined by the subtopics and the categories of activities, six topics are outlined because they represent the combination of an important substantive area and a major scientific or implementation challenge.

#### Panel 1: Key topics required for global-health metrics and evaluation

##### Health outcomes

- Mortality
- Causes of death
- Disease incidence and prevalence
- Functional health status
- Burden of disease
- Comparative risk assessment

##### Other outcomes

- Fairness in financing
- Responsiveness and satisfaction
- Inequalities

##### Health services

- Intervention effective coverage
- Provider technical quality

##### Resource inputs

- Financial resources
- Human resources
- Drugs, equipment, facilities

##### Evaluations

- Interventions and programmes
- National health systems

##### Supporting policy choice

- Forecasting
- Priority setting analysis

#### Functional health status

Death rates do not tell the whole story about population health. For many conditions such as blindness, mental disorders, or musculoskeletal diseases, the main effects are loss of health function. For example, years lived with disability (YLDs) is a commonly used metric summarising functional health status.<sup>13</sup> 37% of disability-adjusted life years (DALYs) worldwide are due to YLDs—a percentage that rises to 56% in North America.<sup>14</sup> Within high-income countries, especially in the context of clinical trials, there is substantial work on measuring functional health status.<sup>15,16</sup> It is measured by a series of core health domains such as vision, hearing, cognition, mobility, dexterity, pain, or affect.<sup>17</sup> Despite innovative efforts to enhance comparability with methods such as anchoring vignettes,<sup>18–20</sup> little progress has been made in the comparable measurement of functional health status.<sup>21</sup> Development of instruments that can be used across cultural and linguistic groups will be crucial to the field of metrics and evaluation.

#### Effective coverage

The notion of effective coverage was introduced as part of the refinement of the WHO framework for health systems performance assessment.<sup>22</sup> Effective coverage is the part of potential health gain that is delivered to a population. It advances the idea of coverage—namely, use of an intervention conditional on need, by incorporation of the quality of the intervention delivered.

Effective coverage has been used to benchmark health-system performance at the state level in Mexico.<sup>23</sup> The challenge for broader implementation of this approach is both to expand the set of interventions that can be tracked beyond the health-related Millennium Development Goals (MDGs) and also to develop methods and data systems to capture the quality of interventions. A promising strategy would be to use biomarkers that can identify when an intervention has been delivered effectively and can be implemented in field conditions, such as tetanus toxoid antibodies for DPT3 delivery, viral load for antiretroviral delivery, or haemoglobin A<sub>1c</sub> for glycaemic control.<sup>24</sup> Prospective registries of patient outcomes such as the cohort system implemented for tracking success in directly observed treatment (DOTS) can also be a useful strategy to pursue for other interventions.

### Human resources

In the past 3 years, the importance of expansion of the quantity and quality of human resources in health systems of low-income and middle-income countries has been widely recognised.<sup>25,26</sup> This is a complex field at the intersection of labour economics and health policy.<sup>27,28</sup> Despite substantial efforts by WHO, national data for human resources have many restrictions<sup>26</sup>—eg, efforts to measure migration are available for only a few countries and depend on data systems of the recipient country.<sup>29</sup> Many categories of health workers such as doctors, nurses, or midwives are not comparable across countries in terms of training or experience.<sup>30</sup> Professional registry or licence data are often used but are not always updated by capturing information about migration or death of health workers. Comprehensive census microdata, which can potentially provide a better snapshot of the health workforce, are often not available to analysts in the health sector.<sup>31</sup> The time has come to prioritise the measurement of human-resource quantity and quality across countries.

### Impact evaluations

Evaluation of health improvement due to the adoption of specific policies and programmes is crucial for global health.<sup>5,6</sup> Three inter-related but distinct issues need to be addressed through methodological research and the adoption of norms and standards: (1) identification of the true quantity of interest in impact evaluation; (2) dealing with unmeasured confounding; and (3) addressing the limited overlap in the characteristics of people receiving and those not receiving an intervention.

In economics and other social sciences,<sup>32,33</sup> effect size is usually defined as the average treatment effect if the treatment were applied to the entire population, although some studies do report the average treatment effect in patients who actually received the treatment. Efficacy studies on drugs, vaccines, and procedures attempt to answer the question: what is the effect in a select subset

of individuals usually with no comorbidities, who are encouraged to have the best possible compliance and receive the intervention under conditions of ideal provider behaviour?<sup>6</sup> However, for most policy consideration, the relevant quantity of interest is the expected effect in the groups that need or are likely to receive an intervention, given the providers who will actually deliver it. This notion has been frequently referred to in the biomedical field as effectiveness. A major problem arises when the methods used to deal with confounding, such as randomisation, are mixed up with the definition of the quantity of interest that is being estimated.

Many fields, including economics, sociology, ecology, and public health, have been advancing methods to address the challenge of unmeasured confounding.<sup>32,34</sup> The most robust method for this purpose is randomisation. The increasing use of this method to study the average treatment effect in those who will probably receive an intervention, including interventions delivered to entire communities, shows that it can be applied in many contexts.<sup>35–39</sup> Because of political acceptability, the level of the health-care system at which the intervention operates, or the ethical principle of equipoise,<sup>40</sup> randomisation might not be possible. Dealing with unmeasured confounding in non-randomised evaluations is fundamentally difficult; the range of approaches and methods to be used in these settings needs further development and creation of norms and standards.<sup>41</sup>

A separate but fundamental difficulty is the extent to which one should go beyond the available data. On the one hand, King and Zeng<sup>42</sup> make a powerful argument that treatment effects should not be predicted in groups

#### Panel 2: Activities for the advancement of global-health metrics and evaluation

- Development of new statistical methods and easy-to-use software to aid their application
- Development of software and hardware to facilitate data collection through administrative records, surveys, or censuses
- Setting global norms and standards for data capture, reporting and transmission such as the International Classification of Diseases (ICD)<sup>40</sup> and the System of Health Accounts (SHA)<sup>41</sup>
- Increasing the availability of high-quality primary data.
- Systematic analysis and synthesis of existing datasets to inform the public, research community, and decision makers
- Strengthening the capacity to obtain, analyse, and use data through efforts such as those coordinated by the Health Metrics Network<sup>42</sup>
- Reporting and disseminating results to broad technical and non-technical audiences

that differ in terms of their social, economic, and health-system characteristics from the zone of overlapping covariates in treated individuals and controls. On the other hand, informing real policy choice very often requires estimation of the probable treatment effect in individuals and groups that are very different in characteristics from those included in randomised trials or non-randomised evaluations. In view of the huge potential for these results to be sensitive to the statistical model used to predict a treatment effect, methods need to be developed and norms and standards established to bring about honest assessments of uncertainty in the likely treatment effect.

For Google Scholar see  
<http://scholar.google.com>

### Health-system evaluations

*The World Health Report 2000*,<sup>43,44</sup> which presented a conceptual framework for health systems and applied it to rank countries' performance, catalysed a major expansion of work in this subfield.<sup>45-47</sup> Improved methods and better data<sup>48-50</sup> have increased the opportunities for evaluating health systems. Rigorous country-specific evaluations, such as those undertaken for Mexico,<sup>51</sup> provide important insights into the complexities of reforming health systems. The Health Care Quality Indicators Project by the Organisation for Economic Co-operation and Development (OECD)<sup>52</sup> shows broad interest in improved metrics for health systems. We believe that it will be increasingly important to bring together work on resource inputs to public health and medical care, effective coverage of key interventions, and health outcomes to identify which systems achieve the greatest health gains and which produce units of health gain at the lowest cost. Ultimately, improved understanding of the levels and determinants of health-system performance and efficiency has the potential to lead to shared learning across systems that can benefit all countries.

### Country health information systems

To increase the availability of high-quality primary data, national capacities need to be strengthened to gather data, including investments in software and hardware platforms as well as changing the culture around release of public data. Furthermore, specific methods need to be developed or improved to ease data collection. In many countries, a tension exists between the need to obtain valid and reliable data, often at high cost, and the need for timely local information. In practice, surveys are often used to provide valid measurements nationally, whereas local decision makers have to rely on administrative record systems. New methods are needed to improve the validity and reliability of timely local measurements at a reasonable cost. Promising areas for research include the use of lower cost sampling methods with larger design effects, record links between surveys and administrative systems allowing estimation of selection bias in administrative systems, and Bayesian methods for local-area estimation.

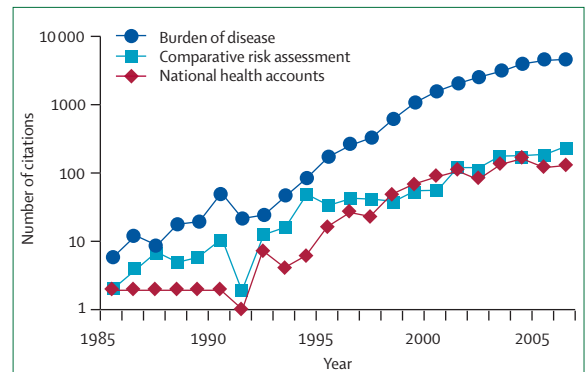


Figure: Yearly number of citations in Google Scholar, 1985–2006

## Institutional and cultural barriers

### Underfunded national institutions

The first of five factors that create institutional barriers to the expansion of the role of scientific measurement and evaluation is underfunded national institutions.<sup>53</sup> What is the root cause of these weak institutions? The evidence for the present state in many countries is not clear. One hypothesis is that a vicious cycle of low demand and poor supply exists. Decision makers might not demand high-quality data, and therefore statistical data systems are underfunded. Because of chronic underfunding, these institutions might then produce low-quality data, which in turn sustain the low demand from decision makers. If this hypothesis is correct, the cycle needs to be broken either by showing decision makers the power of good analyses and thus inducing increased demand, or by external financing of improved supply of high-quality information. The most effective strategy would probably be a combination of both. This vicious cycle implies that in addition to increasing the capacity to gather and analyse data, empowering decision makers to effectively make use of analytical results might be important.

### Institutional entanglement

Globally, one of the main institutional barriers to increasing the role of scientific measurement comes from entanglement. This term<sup>54</sup> captures the situation by which an institution takes on simultaneous roles that are not mutually compatible. For example, institutions that advocate for funding for particular policies, implement programmes, or provide technical assistance should not also be charged with monitoring and evaluating their own progress. The trend towards increasing disentanglement of public institutions nationally<sup>55</sup> has not yet fully worked its way through the international arena. Entanglement can decrease public confidence in statements about trends in global health or attribution of trends to particular policies and programmes. It also inhibits the progressive shift towards more scientific measurement and analysis.

### Low academic status for descriptive analyses

With use of the cases of burden of disease analysis, national health accounts, and comparative risk assessment, the figure shows that in the 1980s, very few descriptive analyses for the levels and trends in global health were published. Few opportunities for publication made the field of metrics and evaluation less attractive to young researchers. However, the fairly low academic status for descriptive studies has been changing. The HIV epidemic has shown their importance for understanding the distribution over time and place of transmission and the use of prevention and antiretroviral treatment. Similarly, the Global Burden of Disease study and the widespread use of DALYs<sup>56–58</sup> in health policy have increased the interest in descriptive analyses. Sustaining and reinforcing the shift towards publication in the leading scientific journals will be essential for the continued expansion of the field.

### Inability to replicate most findings

One of the hallmarks of science is that sceptics can, if they so desire, replicate the results of other scientists by following their published methods.<sup>59</sup> In an era of expensive technology, such as gene sequencers or high-resolution mass spectrometry, the number of scientists who can replicate some findings might be limited. Nevertheless, all experimental findings are replicable by the community as a whole. Although replicability is essential, it is not sufficient to assure scientific progress. Popper<sup>60</sup> states that any scientific theory or research finding should be subjected to the test of falsification. However, this notion also needs transparent reporting of data sources and analytical methods.

To strengthen its scientific basis, global-health metrics and evaluation has to meet the benchmark of replicability. Presently, replicability of most descriptive and analytical epidemiological studies published in the major journals is very low. The underlying data are often not in the public domain, and stated mechanisms to gain access to the data often create powerful barriers for other scientists. Two forces restrict access to primary data and thus to replicability. First, there are legitimate concerns about the protection of individual privacy associated with the public release of some health datasets.<sup>61</sup> The release of microdata into the public domain from more than 150 demographic and health surveys that have been undertaken in more than 70 countries<sup>62</sup> proves that public release is possible when appropriate safeguards are put in place. Second, it is often in the interest of data holders not to release datasets to others. Data holders might not want studies undertaken if they do not control the interpretation, or they might choose to maintain a dataset as a resource for their exclusive analysis and publication. Following the example of the human genome project,<sup>63</sup> it will be crucial to shift the culture in global health towards one of data sharing.

### Perceived risk to decision makers

To support a robust scientific community that undertakes independent measurement and evaluation could be risky for decision makers. Science leads to open debate and critique which can pose political challenges to any decision maker. Yet, at the same time, most decision makers want privately to know how their actions compare with the performance of institutions, communities, or nations in similar circumstances. The challenge is to create an understanding that no decision maker can have the knowledge of how he or she is doing without joining the larger community that is sharing information and fostering scientific debate. Science and democracy have in common the value of promoting open and transparent debate. By embracing this value, decision makers can make an enduring contribution to both scientific and political progress.

### Some scientific challenges and opportunities

Although the landscape of metrics and evaluation is broad, some scientific challenges cut across topics. These challenges are worth drawing attention to since the field will need to address them in a coherent fashion. Everyone in related disciplines is well versed in the difficulties of sampling error, but many of the most daunting challenges relate to non-sampling errors.

### Comparability

The field of metrics pays attention to the challenges of reliability and validity.<sup>15</sup> For measurements with no clear gold standard, different forms of validation have been proposed that are less exacting than criterion validity.<sup>15</sup> A third dimension that is as important is comparability—namely, that a measurement has the same meaning across time and populations.<sup>64</sup> Two thermometers, one in Celsius and one in Fahrenheit, can both give valid and reliable measurements but their results are not comparable. Although putting the results of the two thermometers on a comparable scale is easy, for many global-health measurements the task is much more challenging. Comparability is not simply guaranteed with the same survey instrument or protocol for data collection. The problem whereby the same question on a survey behaves differently depending on the respondents is known as differential item functioning.<sup>65,66</sup> Far more differential item functioning probably exists than is commonly recognised in items on health expenditures, reported illness, use of services, or even socio-demographic characteristics.<sup>18,19</sup> Methods and approaches in psychometrics, survey science, and economics will be needed to study and generate comparability systematically.

### Known bias

Many health measurements have known bias. The number of cases of malaria reported to WHO<sup>67</sup> is a large but variable underestimation across countries;

self-reported obesity in the USA in the behavioural risk factor surveillance survey underestimates obesity in women by 15%.<sup>68</sup> Too often, investigators simply report data and note in the discussion section that the data are known to be biased, which puts the onus of correcting for known bias on the user, who is often less informed and less able to apply the appropriate correction. In contrast, analysts working on national accounts<sup>69</sup> try to correct for bias as do demographers assessing age-specific death rates or fertility.<sup>70</sup> Murray<sup>67</sup> has proposed a distinction between crude statistics, corrected statistics, and predicted statistics. For most substantive purposes, our attention should be focused on the corrected data. Research needs to be undertaken on developing, refining, and applying the appropriate analyses to correct for known bias in a manner that meets the benchmark of replicability.

#### Missing data

Nearly all datasets used for global-health metrics and evaluation have substantial issues with missing data. Some administrative systems such as the antenatal clinic sero-surveillance for HIV in sub-Saharan Africa have 77% of missing data from 1995–2005. In addition to the well characterised bias in estimation of causal effects that missing data can induce,<sup>71</sup> missing data in time-series cross-sectional datasets can lead to the interpretation of spurious time trends. So far, the field of metrics and evaluation does not have standard approaches for addressing the challenge of missing data; some reports do not even mention how missing data have been handled. In view of the widespread nature of this problem, research is needed to identify the best possible methods to deal with missing data for the task of tracking trends in global health and for evaluative studies.

#### Full uncertainty analysis

All major journals require that uncertainty be reported in key quantities of interest because of sampling error, but often do not require capturing uncertainty due to other sources. Generally, uncertainty comes from many sources. Many different models can be used to address the same question, and the set of all plausible models will generate a wider range of results than will one model. For any given model, sampling error in datasets means that the estimated parameters are also uncertain, which translates into uncertainty in the results or predictions of models. Beyond uncertainty in the parameters, there is often much unexplained variation in the analysis, which should also be shown in the final results. With the substantial increase in affordable computing power, numerical simulation methods can be used to adequately assess the uncertainty in a quantity of interest because of model choice, parameter uncertainty, and unexplained variation.<sup>72</sup> Nevertheless, few global-health tracking studies report full uncertainty assessments. For example, uncertainty estimates for child mortality were published

only in 2007.<sup>73</sup> The estimation and reporting of full uncertainty needs much improvement in the available methods and standardisation of reporting requirements for journals and other scientific publications.

#### Validation of disease models

Mathematical models of disease progression or transmission, or both, have had important roles in cost-effectiveness studies predicting the effect of new policies or interventions<sup>74</sup> and in the interpretation of epidemiological data such as antenatal clinic sero-surveillance.<sup>75–77</sup> Development of a disease model entails many judgments, and scientists can and do develop alternative simplifications.<sup>78,79</sup> Most published analyses, however, only use one of a large selection of models in making their conclusions. Moreover, the many parameters in these models are most often not estimated by the standard methods of statistical inference, but are based on selective review of published work, parameter by parameter. Although analysts<sup>80</sup> do calibrate their models to some observations from real data, this practice is usually not based on standard methods of statistical inference.

Why are disease models not subject to the same evidence standard as all other global-health analyses? The usual argument is that there are not enough observed data to fit models to with use of the methods of statistical inference. If formal statistical inference is not possible, the model parameters cannot be known with a reasonable certainty. Logically, the uncertainty in model predictions should be very large, yet most published models generate quite narrow uncertainty intervals. The difficulty of validating disease models has drawn a lot of attention.<sup>81,82</sup>

Establishing disease modelling on a more sound scientific footing requires a two-pronged approach. First, the task for estimation of disease-model parameters should be firmly based on the established methods of statistical inference. Biological or physical knowledge about parameter values can be captured in this framework with Bayesian methods. Second, conclusions or recommendations should never be based on one disease model specification, but rather on an extensive set of competitive models that purport to be adequate representations of a disease progression or transmission process.

#### Retrospective and prospective analysis of cost-effectiveness

Disease models are combined with cost estimates to generate cost-effectiveness results that are meant to inform policy formulation.<sup>83,84</sup> Concerns about the dependence of cost-effectiveness analysis (CEA) results on a wide range of investigator assumptions have led to the idea in high-income countries that costs and computation of cost-effectiveness should be embedded in clinical trials.<sup>85,86</sup> This effort ensures more accurate

assessments of cost efficacy but does not necessarily provide realistic evaluations of the cost-effectiveness of interventions that are implemented in settings outside of clinical trials. In global health, most cost-effectiveness analyses are undertaken prospectively—ie, before a policy or intervention has been applied in real populations. How realistic are these prospective assessments? Substantial differences in published prospective analyses of the same intervention, such as for improved stoves to reduce indoor air pollution in southeast Asia—which range from US\$14 per DALY averted to \$897 per DALY averted<sup>74,87</sup>—emphasise the need for more validation. Prospective cost-effectiveness studies can ultimately only be truly validated by systematic comparisons to the cost-effectiveness of the same intervention when applied in a real population—ie, by assessments undertaken retrospectively. As with the use of funnel plots to detect publication bias,<sup>88</sup> a simple test of validity for ex-ante cost-effectiveness analyses would be to plot the prospective cost-effectiveness ratios against the retrospective ratios. We should expect that 95% of the time the retrospective results will be within the 95% uncertainty interval of the prospective studies. However, at present there are simply not enough retrospective CEAs in global health to study if this notion is true. We strongly suspect that prospective assessments are on average much more optimistic than retrospective assessments because of the natural tendency to underestimate the effect of low household demand, low adherence, and poor provider performance on the costs and effects of an intervention. Real progress has been made in the systematic assessment of intervention costs and effectiveness. The next step in strengthening this evidence base will be to use retrospective and prospective comparisons.

### Building an interdisciplinary community for the future

The foregoing challenges and opportunities refer to the production of knowledge around health metrics and evaluation. To build a vigorous research tradition requires first to define the boundaries of the field and to develop institutions where researchers can generate theoretical formulations, methodological standards, and a body of substantive discoveries.

As important as the production of knowledge is its reproduction, which contributes to the consolidation and continuity of the field of enquiry, and is achieved through three major means: first, publications to assure that the new knowledge reaches a wide audience and is subject to peer scrutiny; second, educational programmes to train the next generations of researchers and practitioners; and third, scientific and professional associations to foster the exchange of ideas and to forge a sense of community.<sup>89</sup>

In the case of health metrics and evaluation, dedicated publications are only beginning to emerge. A noteworthy

effort is the Global Health Tracking initiative, which is launched in this issue of *The Lancet* as a collaboration with the Institute for Health Metrics and Evaluation. Hopefully other journals will also expand the dissemination of rigorous measurements and evaluations related to global health.

With respect to educational programmes, the key challenge will be to train analysts and researchers that embody, in their own competence set, the interdisciplinary nature of the field of health metrics and evaluation—ie, hybrid types of analysts who can span methodological and substantive aspects of the field, while mastering the knowledge, skills, and attitudes required to engage in team work across many disciplines, methods, and content areas.

The final means to assure the reproduction of a field is the development of associations of researchers and practitioners. As the number of people working on health metrics and evaluation continues to grow, an academy or a similar group that can serve as a forum for the advancement of the field will need to be established.

In addition to the production and the reproduction of knowledge, mechanisms need to be developed for its use by policy makers, programme managers, frontline practitioners, and the public at large. For this reason, the field of metrics and evaluation has to combine three fundamental values: excellence in the strict adherence to the highest standards of scientific rigour, relevance to decision making, and independence from any particular interest.<sup>90</sup>

We are entering a new era in global health, which is characterised by unprecedented prominence, but also by expansion of expectations. The only way to sustain such prominence and to respond to these expectations will be to inform purposeful action with evidence derived from sound metrics and evaluation. In our complex and contradictory world, science remains the guiding force for enlightened social transformation. To prosper, the scientific enterprise needs strong institutions that can facilitate the production, reproduction, and use of knowledge. Solid metrics and evaluation will be an essential factor to ensure that the new era of global health delivers its full potential for improving the wellbeing of populations.

#### Conflict of interest statement

We declare that we have no conflict of interest.

#### Acknowledgments

We thank Stanislava Nikolova for editorial assistance.

#### References

- 1 Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; **312**: 71–72.
- 2 Kohatsu ND, Robinson JG, Torner JC. Evidence-based public health: an evolving concept. *Am J Prev Med* 2004; **27**: 417–21.
- 3 Friedrichsen GWS, Burchfield RW, Onions CT. *The Oxford Dictionary of English Etymology*. Oxford: Oxford University Press, 1966.
- 4 Szreter S. The GRO and the public health movement in Britain, 1837–1914. *Soc Hist Med* 1991; **4**: 435–63.

- 5 Buekens P, Keusch G, Belizan J, Bhutta ZA. Evidence-based global health. *JAMA* 2004; **291**: 2639–41.
- 6 Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. *Am J Public Health* 2004; **94**: 400–05.
- 7 Kirkwood B. Making public health interventions more evidence based. *BMJ* 2004; **328**: 966–67.
- 8 The Lancet. A new institute for global health evaluations. *Lancet* 2007; **369**: 1902.
- 9 Madon T, Hofman KJ, Kupfer L, Glass RI. Public health: implementation science. *Science* 2007; **318**: 1728–29.
- 10 WHO. International statistical classification of diseases and related health problems, 10th revision, version for 2007. Geneva: World Health Organization. <http://www.who.int/classifications/icd/en/> (accessed Feb 25, 2008).
- 11 Organisation for Economic Co-operation and Development. A system of health accounts. Paris, OECD, 2000. <http://www.oecd.org/dataoecd/41/4/1841456.pdf> (accessed Feb 25, 2008).
- 12 WHO. Health Metrics Network. Geneva: World Health Organization. <http://www.who.int/healthmetrics/en/> (accessed Feb 25, 2008).
- 13 Murray CJL, Lopez AD. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: results of the Global Burden of Disease Study. *Lancet* 1997; **349**: 1347–52.
- 14 WHO. Revised global burden of disease (GBD) 2002 estimates. Geneva: World Health Organization, 2002. <http://www.who.int/healthinfo/bodgbd2002revised/en/index.html> (accessed Feb 25, 2008).
- 15 McDowell I. Measuring health: a guide to rating scales and questionnaires (3rd ed). Oxford: Oxford University Press, 2006.
- 16 Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Int Med* 1993; **118**: 622–29.
- 17 Salomon JA, Mathers CD, Chatterji S, Sadana R, Ustun TB, Murray CJL. Quantifying individual levels of health: definitions, concepts, and measurement issues. In: Murray CJL, Evans DB, eds. Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization, 2003: 301–18.
- 18 Salomon JA, Tandon A, Murray CJL, World Health Survey Pilot Study Collaborating Group. Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *BMJ* 2004; **328**: 258.
- 19 King G, Murray CJL, Salomon JA, Tandon A. Enhancing the validity and cross-cultural comparability of measurement in survey research. *Am Polit Sci Rev* 2003; **97**: 567–84. Reprinted, with printing errors corrected, 2004 Feb; **98**(1): 191–207.
- 20 Kapteyn A, Smith JP, van Soest A. Vignettes and self-reports of work disability in the United States and the Netherlands. *Am Econ Rev* 2007; **97**: 461–73.
- 21 Murray CJL, Salomon JA, Mathers CD, Lopez AD, eds. Summary measures of population health: concepts, ethics, measurement and applications. Geneva: World Health Organization, 2002.
- 22 Shengelia B, Tandon A, Adams OB, Murray CJ. Access, utilization, quality, and effective coverage: an integrated conceptual framework and measurement strategy. *Soc Sci Med* 2005; **61**: 97–109.
- 23 Lozano R, Soliz P, Gakidou E, et al. Benchmarking of performance of Mexican states with effective coverage. *Lancet* 2006; **368**: 1729–41.
- 24 Sgaier SK, Jha P, Mony P, et al. Public health. Biobanks in developing countries: needs and feasibility. *Science* 2007; **318**: 1074–75.
- 25 Chen L, Evans T, Anand S, et al. Human resources for health: overcoming the crisis. *Lancet* 2004; **364**: 1984–90.
- 26 WHO. The world health report 2006—working together for health. Geneva: World Health Organization, 2006.
- 27 Vujcic M, Zurn P. The dynamics of the health labour market. *Int J Health Plann Mgmt* 2006; **21**: 101–15.
- 28 Pond B, McPake B. The health migration crisis: the role of four Organisation for Economic Cooperation and Development countries. *Lancet* 2006; **367**: 1448–55.
- 29 Mullan F. The metrics of the physician brain drain. *N Engl J Med* 2005; **353**: 1810–18.
- 30 Diallo K, Zurn P, Gupta N, Dal Poz M. Monitoring and evaluation of human resources for health: an international perspective. *Hum Resour Health* 2003; **1**: 3.
- 31 Gupta N, Zurn P, Diallo K, Dal Poz MR. Uses of population census data for monitoring geographical imbalance in the health workforce: snapshots from three developing countries. *Int J Equity Health* 2003; **2**: 11.
- 32 Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. National Bureau of Economic Research, October 2005. <http://www.nber.org/confer/2005/lsof05/imbens.pdf> (accessed Feb 25, 2008).
- 33 Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–60.
- 34 Carpenter SR, Frost TM, Heisey D, Kratz TK. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. *Ecology* 1989; **70**: 1142–52.
- 35 Morris SS, Flores R, Olinto P, Medina JM. Monetary incentives in primary health care and effects on use and coverage of preventive health care interventions in rural Honduras: cluster randomised trial. *Lancet* 2004; **364**: 2030–37.
- 36 The Gambia Hepatitis Intervention Study. The Gambia Hepatitis Study Group. *Cancer Res* 1987; **47**: 5782–87.
- 37 Miguel E, Kremer M. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004; **72**: 159–217.
- 38 Kidane G, Morrow RH. Teaching mothers to provide home treatment of malaria in Tigray, Ethiopia: a randomised trial. *Lancet* 2000; **356**: 550–55.
- 39 Rivera JA, Sotres-Alvarez D, Habicht J-P, Shamah T, Villalpando S. Impact of the Mexican Program for Education, Health, and Nutrition (Progresa) on rates of growth and anemia in infants and young children: a randomized effectiveness study. *JAMA* 2004; **291**: 2563–70.
- 40 Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987; **317**: 141–45.
- 41 Ravishankar N, Lim S, Obermeyer Z, Murray CJL, Gakidou E. Doris Duke Charitable Foundation. PHIT Partnership Implementation Research Framework. Doris Duke Charitable Foundation, 2008. [http://www.ddcf.org/doris\\_duke\\_files/download\\_files/080225DDCFIRFramework.pdf](http://www.ddcf.org/doris_duke_files/download_files/080225DDCFIRFramework.pdf) (accessed Feb 25, 2008).
- 42 King G, Zeng L. When can history be our guide? The pitfalls of counterfactual inference. *Int Stud Q* 2007; **51**: 183–210.
- 43 Murray CJL, Frenk J. A framework for assessing the performance of health systems. *Bull World Health Organ* 2000; **78**: 717–31.
- 44 WHO. The world health report 2000—health systems: improving performance. Geneva: World Health Organization, 2000.
- 45 Murray CJL, Evans DB, eds. Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization, 2003.
- 46 Organisation for Economic Co-operation and Development, ed. Measuring up: improving health system performance in OECD countries. Paris: OECD, 2002.
- 47 Kawabata K. A new look at health systems. *Bull World Health Organ* 2000; **78**: 716.
- 48 Jacobs R, Smith PC, Street A. Measuring efficiency in health care. Analytic techniques and health policy. New York: Cambridge University Press, 2006.
- 49 Greene W. Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Econ* 2004; **13**: 959–80.
- 50 Evans DB, Tandon A, Murray CJ, Lauer JA. Comparative efficiency of national health systems: cross national econometric analysis. *BMJ* 2001; **323**: 307–10.
- 51 Gakidou E, Lozano R, González-Pier E, et al. Assessing the effect of the 2001–06 Mexican health reform: an interim report card. *Lancet* 2006; **368**: 1920–35.
- 52 Organisation for Economic Co-operation and Development. Health Care Quality Indicators Project. Paris: OECD. <http://www.oecd.org/health/hcqi> (accessed Feb 25, 2008).
- 53 Boerma JT, Stansfield SK. Health statistics now: are we making the right investments? *Lancet* 2007; **369**: 779–86.
- 54 Rychetnik L, Hawe P, Waters E, Barratt A, Frommer M. A glossary for evidence based public health. *J Epidemiol Community Health* 2004; **58**: 538–45.
- 55 Shaw C. External assessment of health care. *BMJ* 2001; **322**: 851–54.

- 56 Mathers CD, Vos ET, Stevenson CE, Begg SJ. The burden of disease and injury in Australia. *Bull World Health Organ* 2001; **79**: 1076–84.
- 57 WHO. The world health report 2002—reducing risks, promoting healthy life. Geneva: World Health Organization, 2002.
- 58 Bradshaw D, Groenewald P, Laubscher R, et al. Initial burden of disease estimates for South Africa, 2000. Cape Town: South African Medical Research Council, 2003.
- 59 Stephan PE. The economics of science. *J Econ Lit* 1996; **34**: 1199–235.
- 60 Popper K. The logic of scientific discovery. New York: Basic Books, 1959.
- 61 Buckovich SA, Rippen HE, Rozen MJ. Driving toward guiding principles. A goal for privacy, confidentiality, and security of health information. *J Am Med Inform Assoc* 1999; **6**: 122–33.
- 62 Macro International. Measure Demographic and Health Surveys. <http://www.measuredhs.com/> (accessed Feb 25, 2008).
- 63 Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science* 2003; **300**: 286–90.
- 64 Murray CJL, Tandon A, Salomon JA, Mathers CD, Sadana R. Cross-population comparability of evidence for health policy. In: Murray CJL, Evans DB, eds. Health systems performance assessment: debates, methods and empiricism. Geneva: World Health Organization, 2003: 705–713.
- 65 Holland PW, Wainer H, eds. Differential item functioning. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1993.
- 66 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000; **38** (9 suppl): 1128–42.
- 67 Murray CJL. Towards good practice for health statistics: lessons from the Millennium Development Goal health indicators. *Lancet* 2007; **369**: 862–73.
- 68 Ezzati M, Martin H, Skjold S, Vander Hoorn S, Murray CJ. Trends in national and state-level obesity in the USA after correction for self-report bias: analysis of health surveys. *J R Soc Med* 2006; **99**: 250–57.
- 69 Eisner R. Extended accounts for national income and product. *J Econ Lit* 1988; **26**: 1611–84.
- 70 UN. Manual X: indirect techniques for demographic estimation. New York: United Nations, 1983. [http://www.un.org/esa/population/publications/Manual\\_X/Manual\\_X.htm](http://www.un.org/esa/population/publications/Manual_X/Manual_X.htm) (accessed Feb 25, 2008).
- 71 King G, Honaker J, Joseph A, Scheve K. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *Am Polit Sci Rev* 2001; **95**: 49–69.
- 72 King G, Tomz M, Wittenberg J. Making the most of statistical analyses: improving interpretation and presentation. *Am J Polit Sci* 2000; **44**: 347–61.
- 73 Murray CJL, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. *Lancet* 2007; **370**: 1040–54.
- 74 Jamison DT, Breman JG, Measham AR, et al, eds. Disease control priorities in developing countries (2nd edn). A copublication of The World Bank and Oxford University Press. Washington, DC: The International Bank for Reconstruction and Development/The World Bank, 2006.
- 75 Brookmeyer R, Damiano A. Statistical methods for short-term projections of AIDS incidence. *Stat Med* 1989; **8**: 23–34.
- 76 Chin J, Lwanga SW. Estimation and projection of adult AIDS cases: a simple epidemiological model. *Bull World Health Organ* 1991; **69**: 399–406.
- 77 Salomon JA, Murray CJL. Modelling HIV/AIDS epidemics in sub-Saharan Africa using seroprevalence data from antenatal clinics. *Bull World Health Organ* 2001; **79**: 596–607.
- 78 Dye C, Garnett GP, Sleeman K, Williams BG. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *Lancet* 1998; **352**: 1886–91.
- 79 Murray CJ, Salomon JA. Expanding the WHO tuberculosis control strategy: rethinking the role of active case-finding. *Int J Tuberc Lung Dis* 1998; **2** (suppl 1): S9–15.
- 80 Salomon JA, Weinstein MC, Hammit JK, Goldie SJ. Empirically calibrated model of hepatitis C virus infection in the United States. *Am J Epidemiol* 2002; **156**: 761–73.
- 81 McCabe C, Dixon S. Testing the validity of cost-effectiveness models. *Pharmacoeconomics* 2000; **17**: 501–13.
- 82 Sheldon TA. Problems of using modelling in the economic evaluation of health care. *Health Econ* 1996; **5**: 1–11.
- 83 Laxminarayan R, Mills AJ, Breman JG, et al. Advancement of global health: key messages from the Disease Control Priorities Project. *Lancet* 2006; **367**: 1193–208.
- 84 Adam T, Lim SS, Mehta S, et al. Cost effectiveness analysis of strategies for maternal and neonatal health in developing countries. *BMJ* 2005; **331**: 1107.
- 85 Kassirer JP, Angell M. The journal's policy on cost-effectiveness analyses. *N Engl J Med* 1994; **331**: 669–70.
- 86 Powe NR, Griffiths RI. The clinical-economic trial: promise, problems, and challenges. *Control Clin Trials* 1995; **16**: 377–94.
- 87 WHO. CHOosing Interventions that are Cost Effective (WHO-CHOICE). Geneva, WHO. <http://www.who.int/choice/en/> (accessed Feb 25, 2008).
- 88 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; **315**: 629–34.
- 89 Frenk J. The new public health. *Annu Rev Public Health* 1993; **14**: 469–90.
- 90 Frenk J. Balancing relevance and excellence: organizational responses to link research with decision making. *Soc Sci Med* 1992; **35**: 1397–404.