Accounting for the cost of scaling-up health interventions

Benjamin Johns\textsuperscript{a,*} and Rob Baltussen\textsuperscript{a,b}
On behalf of WHO-CHOICE

\textsuperscript{a}Global Programme on Evidence for Health Policy (GPE/EQC), World Health Organization, Switzerland
\textsuperscript{b}Institute for Medical Technology Assessment (iMTA), Erasmus Medical Centre Rotterdam, The Netherlands

Summary

Recent studies such as the Commission on Macroeconomics and Health have highlighted the need for expanding the coverage of services for HIV/AIDS, malaria, tuberculosis, immunisations and other diseases. In order for policy makers to plan for these changes, they need to analyse the change in costs when interventions are ‘scaled-up’ to cover greater percentages of the population. Previous studies suggest that applying current unit costs to an entire population can misconstrue the true costs of an intervention. This study presents the methodology used in WHO-CHOICE’s generalised cost effectiveness analysis, which includes non-linear cost functions for health centres, transportation and supervision costs, as well as the presence of fixed costs of establishing a health infrastructure. Results show changing marginal costs as predicted by economic theory. Copyright © 2004 John Wiley & Sons, Ltd.

Keywords health; cost-effectiveness; coverage; scaling-up; costs

Introduction

Recent studies such as the Commission on Macroeconomics and Health have highlighted the need for expanding the coverage of services for HIV/AIDS, malaria, tuberculosis, immunisations and other diseases [1]. In order for policy makers to plan for these changes, they need to analyse the change in costs when interventions are ‘scaled-up’ to cover greater percentages of the population. They, for example, may be interested in the cost of extending health interventions to the poorest people in their country, who often live in rural or remote areas previously uncovered [2]. Without a means of determining the costs of expanding health interventions into such areas, policy makers cannot assess the desirability or feasibility of scaling-up interventions.

Previous costing studies have shown that rural health centres have higher cost per patient treated than urban health centres [3,4], and comparisons of EPI programmes imply that very high levels of population coverage increase the cost per vaccine given [5–7]. Additionally, Kumaranayake and Watts show increasing average costs for HIV/AIDS curative and preventive interventions moving from current to 25% coverage level in sub-Saharan Africa [8]. However, little empirical work has been done on the cost of scaling up interventions, because, in general, costing studies have been point-in-time...
measurements that do not include increases in coverage over time [9].

The few theoretical studies on the subject show that average costs, or cost per recipient, can seriously misconstrue the true costs of an intervention and lead to erroneous conclusions on the cost effectiveness of interventions. While the World Development Report 1993 acknowledged that cost-effectiveness ratios would vary with population coverage, it left no methodology for evaluating these differences. For example, they evaluated EPI programmes at only two coverage levels, using a percentage mark-up to account for the difference in cost [2].

Over theorised that factors influencing the cost of expanding coverage included the transportation costs of supervisory trips and drug re-supply, and the recruitment and training of health personnel in remote areas [10]. Further, capital costs may increase in remote areas because the items needed for building have to be shipped greater distances. Thus, he predicted an eventually increasing marginal cost curve, resulting in a U-shaped average cost curve.

The cost component of the Commission on Macroeconomics and Health (CMH) calculated the cost of scaling-up by multiplying an average cost per patient by the predicted future number of patients, and adding a percentage of these costs to account for training and upgrading the overall health system [11–13].

As part of its WHO-CHOICE project, a WHO has undertaken an effort to assess the overall costs and effects of a wide variety of health interventions, including the cost of scaling-up [14]. WHO-CHOICE is performing this assessment for a range of interventions for 14 sub-regions of the world based on the Global Burden of Disease epidemiological patterns [15]. The provision of sub-regional estimates allows interventions to be classified into broad categories that have broad validity across the region – e.g. those that are very cost-effective, those that are cost-effective, and those that are cost-ineffective. Policy makers can then ask if there are good reasons why very cost-effective interventions are not done in their setting, while at the same time cost-ineffective interventions are being done.

As a first step in the process of assessing costs, WHO-CHOICE reviewed the literature on the cost of scaling-up [Johns B, Tan Torres T. Predicting the cost of scaling-up health interventions: a systematic review. Submitted manuscript, 2003]. This study identifies a number of broad areas where economies or diseconomies of scale exist in health interventions: geography and transportation, fixed costs of establishing a health infrastructure, human resources, and management transition costs. This study also finds that there has been no method developed for accounting for these components of scaling-up in different settings and/or for different types of health interventions. Without taking into account these factors, any attempt to assess the cost of scaling-up health interventions could misconstrue the true costs of these interventions. Since each of the above factors interact differently with different interventions, failing to account for these factors could affect the relative cost-effectiveness of interventions at different levels of coverage.

The objective of this paper is to describe a methodology to estimate the cost of scaling-up for use in cost-effectiveness in general and for WHO-CHOICE’s generalised cost effectiveness analysis (GCEA) in particular. It defines a methodology for accounting for the spatial aspects of scaling-up based on the literature review.

Methods

The assumptions used in WHO-CHOICE’s GCEA affect what and how costs are counted. Importantly, GCEA analysis ask which interventions potentially are the most cost effective, and therefore assumes a standard capacity utilisation rate of 80% [16]. This prevents an (unfair) comparison of the cost-effectiveness of programmes that are currently run very inefficiently to those that are run efficiently. Thus, while the total costs of programme management are included, costs to improve management techniques are not included.

Four issues are distinguished in the accounting the costs of scaling up. Firstly, as mentioned above, several studies suggest that rural health centres have higher cost per patient due to lower capacity utilisation (discussed in Health Centre Unit Costs). Second, scaling-up health care programmes into more remote areas also involves shipping supplies and equipment over longer distances and through more difficult terrain. Fielder shows, for example, that the cost of shipping a kilogram of vitamin A capsules to different provinces in Nepal can have an almost six-fold difference in cost ([17] analysed in Johns...
and Tan Torres (2003)) (discussed in Transportation Costs). Third, studies also show that administrative regions tend to have fixed cost, in terms of media outreach, training, or basic administration, regardless of the population covered [18–20] (discussed in Fixed Costs). Finally, as programmes are scaled-up, the supervision and management become more complicated, requiring more resources devoted to these activities [11,21] (discussed in Supervision Costs). An example of the results for each section are presented for the AfrDc and AmrAd regions, both to demonstrate the results of the analysis and to suggest how the results differ in developing and developed countries.

1. Health Centre Unit Costs: For equity reasons, it is assumed that there is a health centre accessible within 1 h to all of the covered population [22]. Thus, every person is assumed to be able to access a health centre within 1 h of travel time [23–25]. A unique Geographical Information System component was designed such that the square kilometer with the highest population density in a country was assumed to have a health centre. The population served by this health centre was taken out the grid, and the process re-started. In this system, 1 h travel time assumes access to motor vehicle transport where paved roads exist, otherwise all travel is assumed to occur by foot.

Thus, as coverage expands, the catchment area of each additional health centre tends to have fewer people, indicating that capacity utilisation decreases [26]. Since the overhead costs are distributed among fewer patients, the cost per patient will be higher for these health centres. Capacity utilisation equals the actual number of patients seen divided by the number of patients that would be seen at standard capacity. The actual number of patients seen is calculated as the population of the catchment area multiplied by the number of visits per capita per year [26,27, author consultation with WHO]. The number of patients that would be seen at standard capacity equals the number of health workers multiplied by the number of patients seen per health worker per day [26] multiplied by the number of working days in a year.

For example, the smallest health centre with two health workers in Denmark is assumed to be able to handle 37 patients per day at standard capacity, with 230 working days in a year [27, Adam T, Ebener S, Johns B, Evans DB. Cost of Scaling up Health Interventions at Primary Facilities: Multi-Country Analysis. Forthcoming manuscript, 2003]. The average person is assumed to attend a clinic 6 times per year [27]. Thus, a catchment area with 1500 inhabitants would operate at the standard capacity of 80% \( (((1500*6)/\text{(37*230}})) > 1 \), while a catchment area with 1000 people would operate at a capacity of 71% \( (((1000*6)/(37*230)) = 0.71 \). This model assumes that health centres are not over-utilised, since areas with high population can either increase the size of health facilities or construct new facilities.

Figure 1 shows the results of this analysis for the WHO epidemiological sub region AfrD and AmrA. As can be seen in the figure, in the central urban areas, health centres are used at the target capacity, but as the population coverage expands into rural areas, the capacity utilisation drops dramatically. On the basis of these capacity utilisation data, the cost per patient at various coverage levels is then predicted using econometric models [28, Adam et al., 2003]. Also, the AfrD region sees a decline in capacity utilisation at an earlier level of population coverage, at around 65% coverage, than the AmrA region, which does not see declines in capacity utilisation until almost 90% coverage. Two factors determine this difference: the population of AmrA is more concentrated in urban areas, and it has greater access to roads, thus the size of the catchment areas in rural sections can be geographically larger (and thus include more people) than those in AfrD. The sharp drop in marginal capacity utilisation after these points indicates that covering the entire population would be very expensive; however, since WHO-CHOICE evaluates interventions only up to 95% coverage, this paper does not evaluate at what point trekking clinics or health posts with
occasional staffing would be more appropriate than health centres.

2. Transportation costs: The calculation of the cost of transportation is based on a Limão and Venables study showing the percentage change in price of a good based on the distance it travels, the transportation infrastructure, the average GDP per capita of a country, and other variables relating to the availability of seaports, neighbouring trade partners, etc. [29]. This method is applied only to traded goods, since they need to be shipped throughout a country, while local goods are already present in a place, or have relatively low transport costs [16,30].

To calculate the average distance a good travels, the straight line distance between the centre of the capital city of a country and each of the health centres (as discussed in Health Centre Unit Costs) is calculated. The Limão and Venables price elasticity for distance was then used with this estimated distance to calculate the cost of transportation [31]. While this method does not necessarily mean that higher coverage rates will have higher shipping distances, especially if there are major urban centres outside of the capital city, the results for AfrD and AmrA shown in Figure 2 show that this is the trend, as it is for most regions of analysis. While the differences in the y-intercept shown in Figure 2 for these two regions reflects the differing sizes of countries in the two regions, the similarity in slope suggests that distance travelled will increase similarly (although not identically) in many different regions.

3. Fixed costs: It is assumed that each province has a set level, or fixed, overhead cost associated with the administration of a program. For example, a media outreach campaign using provincial television will require a certain number of broadcasts no matter how many people are in the province. With fixed costs associated with each province, there are diseconomies of scale because each additional province is assumed to contain fewer people. In this model, the most populous province is assumed to be covered first, then the second most populous, and so on. Figure 3 presents the results of this analysis for the AfrD and AmrA regions. The dotted line indicates constant economies of scale (i.e. what would happen if every province covered had the same population), while the solid lines shows the results predicted in this model. The figure shows that the AmrA region will have greater diseconomies of scale than the AfrD region. However, this is not a general trend between developed and developing countries; the results are specific to the population distribution and administrative structure of individual countries.

4. Supervision costs: The distance travelled in a supervision visit also increases with higher coverage. At the national and provincial level, this is calculated as twice the distance that a traded good is transported since supervision visits constitute a round trip. Thus, the number of programme staff involved in supervision activities needs to increase both in proportion to the increased distances covered and to account for...
the increased number of provinces, since each province is assumed to need an equal number of supervision visits.

Results

This section presents findings on the cost of interventions to demonstrate the effects of employing the model. Figure 4 shows the average and marginal costs for an outreach intervention to educate community health workers on nutrition and hygiene for the WHO epidemiological sub region AfrD and AmrA. The costs presented here exclude health centre costs, to illustrate the effects of the use of the fixed costs, transportation and supervision aspects of the model on the total cost. Costs were calculated for each 10% level of population coverage. The figure shows that lowest average cost per capita occurs at around 45% coverage for AfrD, while the sharpest increase in costs occurs between 90% and 100% population coverage. For AmrA the pattern is similar, except that the lowest average cost is at lower coverage levels (at about 30%), and the slope of the marginal cost curve is steeper through the mid-range of coverage. The primary reason for this is that the population of the provinces in AfrD are more equally distributed than those of AmrA; the fixed costs of the program are spread more evenly. This figure is similar to the findings presented in the Over study, which also shows diseconomies of scale starting at about a 30% coverage level [10].

Table 1 shows the total costs per capita for an iron supplementation programme and an iron fortification programme for the WHO epidemiological sub regions AfrD and AmrA. The results are presented for the three geographic coverage levels commonly used for WHO-CHOICE’s analysis. These interventions have two very different cost structures; thus, scaling up is expected to affect the costs in different ways. Iron fortification

![Figure 4. Marginal and average cost for hygiene outreach. In Year 2000 International $ (Health centre costs excluded). Costs represent the amount needed per average year of programme implementation](image)

Table 1. Cost per capita* for two select interventions in year 2000 International $

<table>
<thead>
<tr>
<th>Coverage level (%)</th>
<th>Average cost per capita</th>
<th>Increase in costs/capita per % increase in population covered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Iron supplementation</td>
<td>Iron fortification</td>
</tr>
<tr>
<td>AfrD region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>$ 0.10</td>
<td>$ 0.13</td>
</tr>
<tr>
<td>80</td>
<td>$ 0.24</td>
<td>$ 0.10</td>
</tr>
<tr>
<td>95</td>
<td>$ 0.31</td>
<td>$ 0.09</td>
</tr>
<tr>
<td>AmrA region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>$ 0.28</td>
<td>$ 0.11</td>
</tr>
<tr>
<td>80</td>
<td>$ 0.44</td>
<td>$ 0.09</td>
</tr>
<tr>
<td>95</td>
<td>$ 0.55</td>
<td>$ 0.08</td>
</tr>
</tbody>
</table>

*Costs are average annual discounted programme costs of implementing an intervention during 10 years. Note that a straight comparison of cost per capita across interventions is misleading in deciding whether an intervention is of low cost or more expensive at a population level, because there is variation in the target populations for each of these interventions.
shows economies of scale up through 95% population coverage because the relatively high central costs are spread across more people, with only some flattening of the average cost curve due to increased costs at the district level. Iron supplementation, on the other hand, requires supervision and implementation at the health centre level, and thus shows rising costs as coverage expands. This shows that the model is adaptable to different types of interventions.

Discussion

The methods described in this paper represent a first effort to account for the cost associated with increasing geographic coverage of health interventions. As such, several assumptions were made and are further discussed below. Analysts adapting this method should take into account these assumptions and fit them to their own setting where necessary.

The method assumes goods are shipped from the capital city – as opposed to a port or entry point for traded goods – of the country, which impacts two issues. First, this assumption over-estimates costs for countries with multiple entry points for goods, while for countries with one entry point, this assumption may underestimate costs. The capital city is selected as the best compromise to estimate the average cost of transport at the sub-regional level since it is often the point of central warehousing.

Second, this method uses distance to estimate transport cost. This may not capture the difficulty in accessing certain areas in a country with poor transportation infrastructure, difficult terrain, or both. Also, the data used to calculate the price elasticity of transportation are only available at the country level. Developments in the GIS database indicate that in the future, a more detailed level of analysis will be possible [23]. For example, time of travel studies may become available and replace the variables for distance and the transportation index used in the Limão and Venables study. Alternatively, a different method based more closely on GIS data could be developed.

This method accounts only for spatial factors in the scaling-up of interventions and, in the absence of data, does not account for variation in the price of local goods within a country. While mean salary figures were used in the calculation of total costs, countries’ inability to recruit and retain health care workers in remote areas indicate that current mid-level estimates of wages may in fact underestimate the true costs of successfully scaling up a program. More research is needed in this area.

This model is intended for use in different settings. While the examples in this paper are aggregated up to a regional level, the same methods could be used for country level analysis. As the examples presented show, by using observed data from a country, it accounts for the differences in the spatial distribution of the population, transportation infrastructure, and administrative organisation. Much of this data is available at the country level. GIS data on population distribution and transportation infrastructure are needed to implement this model and are becoming increasingly available. In addition, data on patient utilisation of health centres [27] and the capacity of a minimally sized health facility are needed to model health centre distributions. While this model utilises averages derived from the literature for the latter data [Adam et al., 2003], countries could do their own assessment. Finally, while existing econometric models can be employed to show the increase in average costs for health centres as capacity utilisation falls, countries could also make their own estimates based on observed data in their country.

This method represents a significant improvement in the estimation of the costs of scaling-up interventions. While previous studies scaled cost up linearly, this method accounts for an increase in marginal costs as coverage increases, which fits in with economic theories on economies of scale [15]. Allowing some costs to be fixed regardless of the size of the population reached – for example, television broadcasts – has incorporated economies of scale. On the other hand, using higher cost for transport, supervision, and health centre care has included diseconomies of scale at higher coverage levels. This is an important step for showing the impact of higher coverage on costs and outcomes.

Acknowledgements

The authors thank Chris Murray, David Evans, and Tessa Tan Torres for guidance throughout this project.
We are also grateful to Raymond Hutubessy, Taghreed Adam, Dan Chisholm, and Moses Aikins for their input in the development of the methods used, and to Nataly Sabharwal and Steeve Ebener for their work in processing the data necessary for the development of this method. The comments by two anonymous referees on an earlier draft of this paper are gratefully acknowledged. The views expressed are those of the authors alone and not necessarily those of the organization they represent.

Notes

a. WHO seeks to provide the evidence decision-makers need to set priorities and improve the performance of their health systems. The Global Programme on Evidence for Health Policy (GPE) is assembling regional databases on the costs, impact on population health and cost-effectiveness of key health interventions. This work known as WHO-CHOICE started in 1998 with the development of standard tools and methods. The objectives of WHO-CHOICE are:

- develop a standardised method for cost-effectiveness analysis that can be applied to all interventions in different settings;
- develop and disseminate tools required to assess intervention costs and impacts at the population level;
- determine the costs and effectiveness of a wide range of health interventions, presented with probabilistic uncertainty analysis;
- summarise the results in regional databases that will be available on the Internet;
- assist policy-makers and other stakeholders to interpret and use the evidence.

b. Capacity utilisation is defined as the ratio of productive use of an input over the total potential productive use of an input. For WHO-CHOICE’s purposes, capacity utilisation was assumed to be 80%. However, analyst adapting this method should adjust the capacity utilisation level to that appropriate for their setting.

c. The AfrD region includes Algeria, Angola, Benin, Burkina Faso, Cameroon, Cape Verde, Chad, Comoros, Equatorial Guinea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Madagascar, Mali, Mauritania, Mauritius, Nigeria, Niger, Nigeria, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, and Togo.

d. The AmrA region includes Canada, the United States of America, and Cuba.

e. Health centres are defined as a place for outpatient clinical visits with at least one full time doctor and nurse on the staff.

f. The documentation of this component is in preparation. WHO plans to make this software freely available to the public.

g. Results of this analysis are presented on the WHO-CHOICE website: http://www.who.int/evidence/cea.

h. Note that for certain traded items, such as drugs that require cold chain storage, this model will not capture the full cost of transport. Thus, the costs for distributing these items should be calculated independently.

References


11. Hanson K, Ranson K, Oliveira-Cruz V, Mills A. Constraints to scaling up health interventions: a


