Some problems with international comparisons of health spending

– and a suggestion about how to quantify the size of the problems

Hans Olav Melberg
Department of Health Management and Health Economics,
University of Oslo & HERO

UNIVERSITY OF OSLO
HEALTH ECONOMICS RESEARCH PROGRAMME

Working paper 2011: 4

HERO

Some problems with international comparisons of health spending

– and a suggestion about how to quantify the size of the problems

Hans Olav Melberg
Health Economics Research programme at University of Oslo - HERO
Department of Health Management and Health Economics,
University of Oslo, PO Box 1089, Blindern, N-0317 Oslo, Norway,
e-mail: hans.melberg@gmail.com

Health Economics Research Programme at the University of Oslo
HERO 2011

Keywords: Income inequality, health inequality, socioeconomic status, welfare states, concentration index, health spending per capita, OECD, Norway.

Health Economics Research Programme at the University of Oslo
Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071 (print version.), ISSN 1890-1735 (online), ISBN 978-82-7756-224-7

ABSTRACT

This paper investigates the extent to which conclusions from international comparison of health spending depend on different adjustment methods. The analysis shows, first, that health spending figures differ significantly because of different accounting standards. More specifically, spending on long term nursing care is not treated the same way in all the countries. Next, the rankings differ depending on whether spending is adjusted for purchasing power parity and health specific purchasing power adjustment. Finally, the paper examines the problem of how to quantify the degree to which an adjustment method changes the outcome of a comparison. A rank based approach sum changes in rank and compare this to maximum rank change possible. A share based approach sum changes in each units share of the total. Both approaches create a measure that is bounded between zero and one, but the share based approach also captures changes that do not result in rank differences.

1. Introduction

Countries are often ranked based on health spending per capita. The ranking is sometimes used to argue in favor of more health spending if the country is low on the list or as an indicator of inefficiency when a country is ranked high in spending but low on indicators of health. For instance, Jacobs and Skocpol (2010) use the OECD figures to argue that "other nations get hugely more bang for the bucks than the United States Does." Similarly Cutler and Ly (2011) argue that the gap between spending and outcome in the US compared to other countries "raises the possibility of substantial waste." The OECD spending figures are also used in regressions that explore the relationship between spending and health status across time and countries (See Nixon and Ulmann, 2006, for a review of 38 papers using spending as an independent variable). All this makes the reliability of health spending figures important. If it turns out that the comparison differ because of accounting standards and adjustment methods, then we should be cautious when trying to draw analytical or policy implications from the rankings.

This article will investigate the extent to which spending and rankings vary depending on different adjustment methods based on information from the OECD Health Data. The analysis shows, first, that health spending figures differ significantly because of different accounting standards. More specifically, long term nursing care spending is not treated the same way in all the countries. Some countries classify it as a health expense, others classify it as social spending. Next, the rankings differ depending on whether spending is adjusted for purchasing power parity or not. This is to be expected. There is a reason one uses PPP adjustment – to adjust for price differences - and this reason implies that there should be some change after applying the adjustment. The true problem is that general PPP adjustment is misleading because the conclusions change if we use health specific purchasing power adjustment instead of using the general consumer price index. Finally, the paper examines the problem of how to quantify the degree to which an adjustment method changes the outcome of a comparison.

2. Validity: Are the numbers comparable?

The OECD data on health spending is based on the system of health accounts (SHA). Total spending on health is defined as the sum of all private and public expenditure and investments in a set of sub-categories. One of these sub-categories is spending on long term nursing care (LTNC). This is an important category and for some countries it constitutes about 25% of all health care spending. However, as shown in Table 1, the reported numbers differ greatly between countries. Some report very little spending on long term nursing care. Even countries that are geographically and institutionally similar, such as the Nordic countries, report very different numbers. This raises the suspicion that a major component of total spending is mismeasured.[1] To what extent this is a problem depends on whether the missing spending is simply classified in another sub-category that is included in the total, or whether it is excluded from health spending overall. If countries just differ in terms of which sub-category they use to report their expenses, the total will still be comparable.

[1] This section is inspired by a paper presented by Jes Søgaard at the Nordic Health Economics Study Group (NHESG) in 2009. He pointed out the problem and has provided a detailed analysis (unpublished manuscript).

Data and documentation from the OECD itself indicate that the problem is not just differences in sub-classification among health expenses. The guidelines for how to treat spending on long term nursing care is vague about whether long term care is a social expense or a health expense. This vagueness has led to different practices in different countries. As shown in Table 1, countries which report little spending on long term nursing care, tend to report more spending on long term social care (LTSC). Since spending on long term social care by definition is not health spending, this difference implies that the figures for total health spending are misleading because of differences in accounting standards.

To quantify the importance of the problem, consider the extent to which the Norwegian health spending is above the Nordic average. If one uses the OECD numbers for total health spending (PPP adjusted, 2007), Norway appears to spend 47% more than the other Nordic countries. This is misleading because the Norwegian figures include spending on long term nursing care which other countries categorize as long term social care. To account for this one can either exclude reported LTNC from all countries or one can include both LTNC and LTSC in the total health spending figures for all countries. In both cases the difference between Norway and the other countries drop from 47% to about 30%. This implies that different accounting standards account for more than one third of the observed difference in spending.

Table 1 Health spending in the Nordic countries, 2007

	Norway	Sweden	Danmark	Finland	Iceland	Nordic average	Difference Norway vs. Average
Total health expenditure	4 763	3 323	3 512	2 840	3 319	3 249	47 %
Long term nursing care (LTNC)	1 177	255	718	325	640	485	143 %
Long term social care (LTSC)	0	1 023	0	387	0	353	
Total health expenditure							
... when LTNC is excluded	3 586	3 068	2 794	2 515	2 679	2 764	30 %
... including LTNC and LTSC	4 763	4 346	3 512	3 227	3 319	3601	32 %

Source: OECD Health Data, 2007, PPP adjusted dollars

3. How much does different conversion methods affect the cost comparison?

Even if we ignore measurement problems, cost comparison can be misleading because of biases in the methods used to compare the costs. This section deals with this problem.

To compare costs it is necessary to convert expenditures to a single unit of measurement. It is possible to use natural units such as "work hours" and "number and amount of pharmaceuticals", but it will result in a multi-dimensional measure. If one wants to avoid this, there are several options (see Gerdtham and Jönsson 2000 more more on this). First, expenditures measured in national currencies can be converted to a common unit, such as Euro or USD, using annual exchange rates. The strength of this procedure is its simplicity, the weaknesses include the fact that the exchange rate does not consider price and cost-differences between countries and that the exchange rate may fluctuate too much in the short run to be a useful unit of conversion. Converting the expenditure using purchasing power parity rates adjusts for this by using the price of a standardized basket of goods and services when converting health spending from national currency units. This takes prices into account, but it also includes prices of goods unrelated to the health sector. A health specific purchasing power parity index would be preferable, but this is not widely available since there is no agreement on standardized basked of health goods and prices are not easily available. When using health specific PPP one should also distinguish between methods that adjust for the price of input factors and methods that are based on the price of outputs. For instance, adjusting health expenditures based on differences in the wage levels of health workers in different countries is an example of adjusting for the price of inputs. Adjusting for the total price of outputs, such as a hip-replacement, may give a different result. Altogether this means that there are at least four important and different methods to adjust costs before comparing health spending across countries:

1. Exchange rates
2. Purchasing Power Parity
3. Input based health specific purchasing power parity
4. Output based health specific purchasing power parity

The question is to what extent the measured spending will differ depending on which of these four measures that is used. Moreover, if the results differ one must ask which method that produce the most useful results. This may differ depending on the type of questions one is interested in, so the conclusion need not be that one conversion method is best.

Figure 1 show the size of the differences in health spending in the OECD countries (relative to US spending) depending on whether spending is measured in exchange rate dollars or purchasing power parity dollars. For some countries the difference is dramatic. For instance, the Norwegian spending figures are slightly higher than the US if measured in terms of dollars based on the exchange rate. However, because Norway is an expensive country, if we

use the official PPP adjusted scale Norwegian health spending is only two thirds of the US health spending. In this case the choice of adjustment method is clearly important and not just a minor detail that does not affect the main outcome of the comparison.

Figure 1 Health spending in OECD countries measured as a proportion of US health spending using exchange rate dollars or purchasing power parity dollars

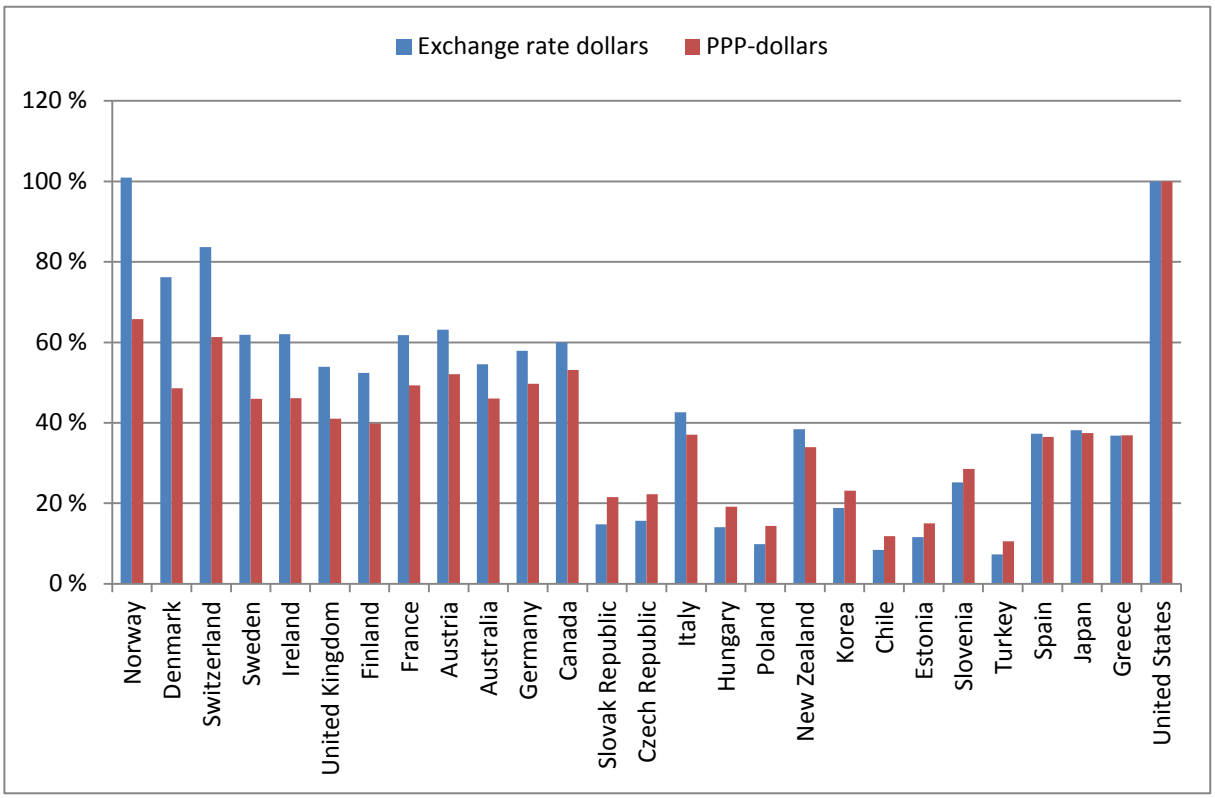


Table 2 Differences in health spending relative to US spending depending on whether the spending is measured in dollars using the exchange rate dollars or purchasing power parity dollars

Country	Difference from US if measured using the exchange rate	Difference from US if measured in PPP-dollars	Absolute difference
Norway	101 %	66 %	35.2 %
Denmark	76 %	49 %	27.6 %
Switzerland	84 %	61 %	22.3 %
Sweden	62 %	46 %	15.9 %
Ireland	62 %	46 %	15.9 %
United Kingdom	54 %	41 %	12.8 %
Finland	52 %	40 %	12.6 %
France	62 %	49 %	12.5 %
Austria	63 %	52 %	11.1 %
Australia	55 %	46 %	8.5 %
Germany	58 %	50 %	8.2 %
Canada	60 %	53 %	6.9 %
Slovak Republic	15 %	22 %	6.8 %
Czech Republic	16 %	22 %	6.6 %
Italy	43 %	37 %	5.6 %
Hungary	14 %	19 %	5.1 %
Poland	10 %	14 %	4.6 %
New Zealand	38 %	34 %	4.5 %
Korea	19 %	23 %	4.3 %
Chile	8 %	12 %	3.4 %
Estonia	12 %	15 %	3.4 %
Slovenia	25 %	29 %	3.3 %
Turkey	7 %	11 %	3.3 %
Spain	37 %	36 %	0.8 %
Japan	38 %	37 %	0.7 %
Greece	37 %	37 %	0.1 %
United States	100 %	100 %	0.0 %

The aim of using PPP dollars as opposed to exchange rate dollars, is to adjust for differences in price (including the price of labor). If the aim is to compare spending in different countries, the relevant price is not the general price level, but the price of health services. This is an important point because some countries, like the US, have relatively high health costs, but low costs for many other goods and services. In this case, using the general price level, as opposed to the health specific price level, when converting to PPP dollars will lead to inflated numbers.

To quantify the importance of the conversion methods, it is necessary to have an index of health prices. There is no such index available on an annual basis from the OECD, but there are some data and previous calculations that can be used. As a first attempt one might use data on wages since wages represent a large share of health expenses. The OECD Health Data contains information about remuneration for nurses and physicians. The information is incomplete and of questionable quality, but it is used here as a starting point to capture differences in input costs.

Information about output prices is even more sparse, and of even worse quality. Once again, as a starting point, it is possible to use some of the results from the HealthBasket project in which expert respondents in nine different countries were asked to calculate typical total costs of treatments based on a brief description of a patient (Busse et al. 2008).[2] The project calculated costs for ten diagnoses, and as a starting point the cost of treating a hip fracture has been used as a proxy of health care prices at the output level (Stargardt 2008). Hip costs were used because information about the cost of treating hip fractures was also available from other countries outside the HealthBasket project.

Using nurse wages as a proxy for input prices and the cost of treating hip fractures as a proxy for output price, Table 3 shows the potential importance of using health specific PPP dollars. Unfortunately, the OECD data on remuneration is very incomplete so the final results are only available for some countries. Once again the conclusion is that the choice of adjustment method matters a great deal. For instance, depending on which measure one chooses, US health spending is 1% less than Norway or 60% higher than higher than in Norway (se Figure 2). To some extent the large difference can be blamed on the incomplete data behind the health PPP, but even in the OECD data one finds differences that are too large to be credible. Even the OECD itself, who has calculated health specific PPP using a wider basis, produced numbers with a "huge difference" depending on whether one used health specific PPP or standard PPP (Castles, 1997). For this reason they continued to rely on standard PPP adjustment, despite its misleading conclusions. For instance, the US has higher health prices than many other countries, but lower prices on other goods and services. In this case, using general PPP inflate the reported US health spending (See Frech 2009 for more on this).

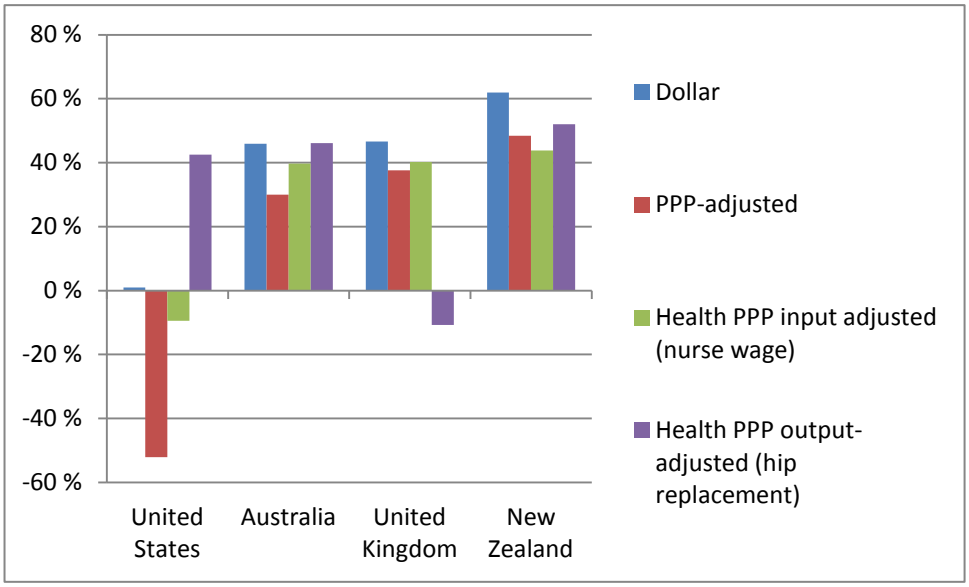
[2] The project and its results is documented in three special issues: Special issue of European Journal of Health Economics, volume 6, November 2005. Special issue of Health Care Management Science, 9(3), 2006 and most importantly (reporting the results): Special issue of Health Economics 17(S1), 2008.

Table 3 Differences between health spending using all four conversion methods

Country	Dollar	PPP-adjusted	Health PPP input adjusted (nurse wage)	Health PPP output-adjusted (hip replacement)
Norway	**7 354**	**4 791**	**6 657**	**12 669**
United States	7 285	7 285	7 285	7 285
Australia	3 975	3 353	4 015	6 824
United Kingdom	3 925	2 990	3 977	14 033
New Zealand	2 798	2 471	3 740	6 078

Source: Calculations based on OECD Health Data, remuneration for nurses and HealthBasket project (Hip treatment costs).

Figure 2: Differences between the Norwegian health expenditures and other countries using different adjustment methods



4. A possible counterargument - and how it could be answered

The conclusion that using health PPP instead of standard PPP produce very different results depends on the choice of which country to use a benchmark. In Figure 2, the differences are calculated using Norway as a benchmark (i.e. the spending of the other countries is measured in terms of deviations from the Norwegian spending). Using a different country as benchmark would produce different results. Moreover, if the benchmark country is more sensitive to changes in adjustment method, then we may get the impression that the method induces large changes when in reality it would not do so for most countries. This

problem applies to all comparisons based on deviations from a benchmark. For instance, the deviation calculated in Table 2 use US spending as a benchmark and using another country as benchmark would produce different results.

To answer the charge that the results depend on the choice of benchmark, it is possible to try out other benchmarks and see if the results differ. An alternative approach would be to develop an index that does not depend on a benchmark. One could, for instance, run the comparison many times using each country as a benchmark and take the overall average as a measure of the degree to which the two methods produced different results. Formally, to compare the difference (Δ) between two adjustment methods, *a* and *b*, on a variable *y* for *n* units (*i=1..n)*, the average in Table 2 correspond to calculating the following using a fixed unit (*j*) as a benchmark (for all $i \neq j$):

$$\Delta_j = \frac{\sum_{i=1}^{n, i \neq j} \left| \frac{y_i^a}{y_j^a} - \frac{y_i^b}{y_j^b} \right|}{n-1}$$

Instead one might calculate the overall average after using each unit as a benchmark:

$$\Delta = \frac{\sum_{j=1}^{j=n} \Delta_j}{n}$$

Although this takes care of the problem of using a benchmark that is misleading, there is a possible problem: There is no upper boundary. This may make it difficult to know whether the end result is small or large.[3] To solve this one could use ordinal rankings. Let *r* symbolize the ranking of a unit among *n* units. We can then take the difference between the ranking of each unit under condition *a* and *b*, and sum the absolute rank differences to get an impression of how much is changed when using a versus b to rank units:

$$\Delta^r = \sum_{i=1}^{i=n} \left| r_i^a - r_i^b \right|$$

This will have an upper boundary since a ranking has a natural maximum - no unit can be ranked higher than 1 or lower than *n*. It is then possible to divide the sum by the maximum possible instability to get a measure where 0 would be no change and 1 would be maximum

[3] The lack of a precise upper boundary need not create a problem. There is no logical upper boundary on a salary and we still know when a salary is high. But in this case it is based on our experience of what a high salary is (a perceived upper boundary).

change.[4] This might be useful sometimes, but often the cardinal information is desired. In the case of health expenditures, the degree to which US spending is higher than the other countries is used as an argument, not only the fact that it has the highest expenditure. Given that there is no upper logical boundary to health spending, it may - at first sight - seem difficult to create an index that respects the cardinality of the data at the same time that it produces a bounded number. However, it is possible to produce a bounded number even if we use the cardinal data. Let s_i symbolize the share unit *i* represent within the sum of all *y*'s

$$s_i = \frac{y_i}{\sum y_i}$$

No single value can be larger than the sum of all values, so this has an upper bound.

Now, for two different adjustment methods, *a* and *b*, we can calculate the difference in the share each unit has of the total depending on whether method *a* or *b* was used:

$$\Delta^s = \sum \left| s_i^a - s_i^b \right|$$

This will have an upper bound since each observation has a maximum (See appendix 2). Unlike the measure based on rankings, this measure is also sensitive to cardinal changes that do not affect rankings.

Calculating the numbers based on the OECD data, produce the following results. When using the US as a benchmark country, and comparing health spending using exchange rates vs. PPP dollars, the average difference in between the two methods was 9.9% Conceptually, this is the average of the absolute differences in Table 2. The ranking difference sums to 47 for 31 observations (average 1.5). Many of the ranking differences are zero, but a few are large. Finally, taking the differences between the shares of the total when spending is measured using exchange rates vs. PPP dollars, gives a sum of 0.17 and an average percentage change (in the share a country represents, depending on whether one measures spending in exchange rate dollars of PPP dollars) of 8.8. More work remains to be done to understand the properties of the various measures, and one should be careful not to focus too much on averages since it may hide large changes among some important units.

5. What should be done?

The fact that international comparisons of health spending is sensitive to accounting differences and adjustment methods is not new. This paper has contributed with some new examples of how large the problem is, but the underlying issues have been known for a long

[4] Interestingly, maximum instability could be achieved when rankings are random i.e. previous ranking gives no information of value for the next ranking, but maximum instability might also be defined by a complete and deterministic reorganization (the last becomes the first and so on). See appendix 1 for more.

time. For instance, Gerdtham and Jönsson described many of the same problems and conducted similar calculations more than 20 years ago. Despite several on- (and long-) going projects by the OECD and the WHO, the problem remains. This raises the question of whether the empirical and conceptual problems are too large to conduct useful international comparisons of health expenditure. The usefulness of a specific adjustment and comparison depends on the aim. If the aim is simply to say something about the volume of health services, then it may be better to stick to natural units in a multidimensional measure. It seems cumbersome and, given the limitations in data and method, unreliable to measure volume in money and then to adjust it for differences in prices in order to get back to a volume measure. Advances in data collection and method may make it more reliable and in some cases perfect precision is not needed to draw conclusions.[5] At the very least one should be aware of the problems. Knowing that the comparison can give almost any result depending on the methods used, is important knowledge in itself since it may vaccinate against unguarded policy conclusions based on an isolated and misleading comparison.

[5] For instance at a WHO conference (Tandon et al. 2002) it has beensuggested that a latent variable approach in which multi-dimensional incomplete health price data can be used to estimate a latent unobserved variable (the general health price level). Although interesting, it seems to assume that the different dimensions of health price are correlated i.e. that high pharmaceutical prices go together with high salaries to nurses.

Appendix 1: The formula for maximum rank difference

The formulas for the maximum possible rank difference sum depends on whether the number of elements (*n*) is even or odd. For even *n*, two different rankings produce a symmetric series of absolute differences. For instance, n = 6, gives the following series of absolute differences when the ranking is reversed: 5, 3, 2, 1, 1, 3, 5 (See illustration below). Ths sum of this series is not immediately obvious, but remembering the probably false story about Gauss' trick for summing the numbers from 1 to 100, we can use a similar trick here. By adding the elements in the lower half to the elements in the upper half, we get a new list of constants: n + n + n. There are n/2 elements on this list, so the sum of all the elements is given by:

$$Max\ rank\ difference\ sum(even\ n) = \frac{1}{2}n^2$$

For odd numbers, things are slightly more complicated because the number of elements on the new list cannot be half that of the original. Instead we have a (n-1)/2 elements each with a value of n+1. The formula for the maximum possible rank difference is then:

$$Max\ rank\ difference\ sum(odd\ n) = \frac{1}{2}(n+1)(n-1)$$

The logic behind the formulas is illustrated in the following table:

Rank A	Rank B	Absolute Difference	Symmetric	Sum
1	6	5	1	6
2	5	3	3	6
3	4	1	5	6
4	3	1		
5	2	3		
6	1	5		

Rank A	Rank B	Absolute Difference	Symmetric	Sum
1	7	6	2	8
2	6	4	4	8
3	5	2	6	8
4	4	0		
5	3	2		
6	2	4		
7	1	6		

Appendix 2: Maximum share difference

For two units, the maximum possible difference is that the first measure or method tells you that unit one makes up the whole of the total while the the other unit makes up all of the total on the other approach. The maximum possible difference is then two. The same result can be generalized. We can go from one unit having everything to another unit having everything in which the absolute difference in shares is 2.

Another approach is to compare an equal distribution with the most unequal distribution possible. Imagine that on one approach all units come out as having an equal share of the total (1/n). Compare this with another approach which indicates that the distribution in most unequal i.e. one unit has everything and all the other units nothing. We then have:

Approach A (equality)	1/n	1/n	…	1/n	1/n
Approach B (max inequality)	0	0		0	1
Absolute difference |A-B|	1/n	1/n	…	1/n	1 - 1/n

The sum of the absolute differences is then

$$\frac{1}{n}+\frac{1}{n}+\ \ldots+\frac{1}{n}+\left(1-\frac{1}{n}\right)$$

$$\frac{1}{n}(n-2)+1$$

$$2-\frac{2}{n}$$

As n approaches infinity, the sum becomes 2, so for large samples, the maximum possible sum of absolute differences in shares is two. For small samples it is still possible to have a difference of two, but in this case it is achieved by going from one having everything to another having everything i.e. the initial distribution is also unequal.

In any case, this establishes an upper bound and enables us to create a standardized measure, between zero and one, of how much of a difference it makes to use one approach compared to the other.

References

Busse, Reinhard, Jonas Schreyögg and Peter C. Smith (2008) Variability in healthcare treatment costs amongst nine EU countries – results from the HealthBASKET project 17(S1):1–8.

Castles, Ian (1997) Review of the OECD-Eurostat PPP Program", unpublished OECD document, STP/PPP(97):5

William Comanor, H.E. French, and Richard Miller, Is the United States an outlier in health care and health outcomes? A preliminary analysis, International Journal of Health Care Finance and Economics, 6(1): 3-23 (2006)

Cutler, David M, Dan P. Ly (2011) The (paper) work of medicine: Understanding international medical costs, Journal of Economic Perspectives 25(2): 3-25.

Frech, H.E. (2009) The OECD's study on health status determinants, AEI Working Paper #145.

Garber, Alan M. and Jonathan Skinner (2008) Is American health care uniquely inefficient? Journal of Economic Perspectives 22(4):27-50.

Gerdtham, U-G, Jönsson, B. Conversion Factor Instability in International Comparisons of Health Care Expenditure. Journal of Health Economics 1991;10: 227-234.

Gerdtham, U-G, Jönsson, B. Price and Quantity in International Comparisons of Health Care Expenditure. Applied Economics 1991;23: 1519-1528.

Gerdtham, U-G, Jönsson, B. Health Care Expenditure in the Nordic Countries. Health Policy 1994; 26: 207-220.

Gerdtham, U-G, Jönsson, B. International Comparisons of Health Expenditure: Theory, Data and Econometric Analysis. In Newhouse, JP, Culyer, AJ (eds.). Handbook of Health Economics. Amsterdam: North Holland/Elsevier, 2000: 11-53.

Gerdtham, U-G, Löthgren, M. On Stationarity and Cointegration of International Health Expenditure and GDP. Journal of Health Economics 2000; 19: 461-475.

Jacobs, Lawrence R., Theda Skocpol (2010) Health care reform and American politics. Oxford: Oxford University Press.

Nixon, John, Philippe Ulmann (2006) The relationship between health care expenditures and health outcomes, European Journal of Health Economics 7(1): 7-18.

Stargardt, Tom (2008) Health service costs in Europe: cost and reimbursement of primary hip replacement in nine countries 17(1):9–20.

Søgaard, Jes (2008) International Reliability of SHA Total Expenditure on Health. Paper presented to the Nordic Health Economics Study Group (NHESG) in Iceland.

Tandon A., J. Klavus, K. Kawabata, D. Evans, and C.J.L. Murray (2002) Cross-population comparability and ppps: Issues relating to health prices, World Health Organization, Conference on the International Comparison Program, World Bank, Washington, D.C., March 11-14, 2002.