OECD Statistics Working Papers 2010/02

Towards Measuring the Volume Output of Education and Health Services: A Handbook

Paul Schreyer

https://dx.doi.org/10.1787/5kmd34g1zk9x-en

OECD

Unclassified **STD/DOC(2010)2**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **28-Apr-2010**

English - Or. English

STATISTICS DIRECTORATE

STD/DOC(2010)2
Unclassified

TOWARDS MEASURING THE VOLUME OUTPUT OF EDUCATION AND HEALTH SERVICES: A HANDBOOK

STATISTICS DIRECTORATE

WORKING PAPER No. 31

JT03282602

English - Or. English

OECD STATISTICS WORKING PAPER SERIES

The OECD Statistics Working Paper Series - managed by the OECD Statistics Directorate - is designed to make available in a timely fashion and to a wider readership selected studies prepared by OECD staff or by outside consultants working on OECD projects. The papers included are of a technical, methodological or statistical policy nature and relate to statistical work relevant to the Organisation. The Working Papers are generally available only in their original language - English or French - with a summary in the other.

Comments on the papers are welcome and should be communicated to the authors or to the OECD Statistics Directorate, 2 rue André Pascal, 75775 Paris Cedex 16, France.

The opinions expressed in these papers are the sole responsibility of the authors and do not necessarily reflect those of the OECD or of the governments of its Member countries.

http://www.oecd.org/std/research

TOWARDS MEASURING THE VOLUME OUTPUT OF EDUCATION AND HEALTH SERVICES

Paul Schreyer is the main author of this document which has also benefitted from contributions by Alain Gallais, Sandra Hopkins, Francette Koechlin, Luca Lorenzoni and Seppo Varjonen.

ABSTRACT

The measurement of volumes of health and education services constitutes a challenge for national accountants and price statisticians. In the past, such services have typically been measured by the inputs used to provide them but such an approach neglects any productivity changes in service provision. An increasing number of countries is now working towards output-based measures of the volume of these services. The present document summarises country practices and provides methodological guidance for output-based approaches in the measurement of health and education services. The handbook deals with volume changes over time within a country as well as with volume differences at a particular point in time across countries.

RESUMÉ

La mesure des volumes de services de la santé et de l'éducation constitue un défi pour les comptables nationaux et les statisticiens des prix. Dans le passé, de tels services ont été typiquement mesurés par les entrants employées pour les fournir mais une telle approche néglige tous les changements de productivité dans la production de service. Un nombre croissant de pays travaille maintenant vers des mesures basées sur une notion de 'output' de ces services. Le présent document récapitule des pratiques en matière de pays et fournit des conseils méthodologiques pour des approches 'output' dans la mesure des services de santé et d'éducation. Le manuel traite des changements de volume temporels au sein d'un pays aussi bien que les différences de volume à un moment particulier à travers des pays.

TABLE OF CONTENTS

Tables

Figures

Boxes

EXECUTIVE SUMMARY

Health and education services have traditionally been measured by inputs, not outputs…

Health and education services account for a sizeable and growing share of GDP in OECD countries. Changes in the volume of these services, in particular when provided by government, have typically been measured by the volume change of inputs such as labour or material rather than by independent measures of outputs. This mainly reflects the difficulties inherent in measuring the output of services produced by non-market producers, where market prices and hence price indices used for deflation cannot be observed, especially when the quality of these services varies over time. In recent years however, several national statistical offices have undertaken work to develop explicit output-based measures. In the European Union, for instance, work has been triggered by statistical legislation that foresees the use of output-based measures in national accounts.

…but outputs are not identical to outcomes

Against this background, this handbook aims at providing guidance for measurement. It also provides information on country practices. The handbook builds on earlier work, in particular by the European Union, but with two special added values. First, discussions have involved national accountants as well as sector specialists from health and education. Second, the handbook considers both the temporal dimension (analysis for a particular country over time) as well as the spatial dimension (analysis across countries at a particular point in time) of volume measurement. The following paragraphs summarize the main conceptual and measurement conclusions.

An important conceptual distinction is that between ***outputs*** and ***outcomes***. Outputs are goods and services that directly result from the production process undertaken in economic units such as schools and health care providers; outcomes are situations that consumers value, such as good health or a high level of education, and these may or may not be a result of production processes as understood by national accountants. National accounts deal with measures of outputs, not outcomes.

As a first approximation, outputs can be captured by observed ***processes*** or activities such as the number of hours students are taught or the number of treatments for patients. Indeed, when there is no quality change in teaching or in treatments, these measures accurately reflect output.

But quality in the provision of services often changes. And so, for education services (provided by schools, universities etc.), the handbook identifies the ***quality-adjusted quantity of teaching*** provided as the basic unit of service.

For health services (provided by hospitals, general practitioners etc.), the handbook identifies as the basic unit of service the ***quality-adjusted numbers of completed treatments of particular diseases or of activities to prevent a disease***.

Services are aggregated with cost weights, not by their value to society

Hours taught at different levels of education are not the same type of service, and cannot therefore simply be added up. Nor are treatments of different diseases the same type of services. To take account of such differences, weights have to be applied in the construction of volume measures. In a market environment, market prices provide these weights. For non-market producers, unit costs can replace prices to value different kinds of services. However, unlike market prices that combine consumer and producer valuations of products, unit ***cost weights*** reflect in the first instance the producer or supply side (or government's willingness to pay). This implies that it is the production value and not necessarily the societal value that is attributed to education or health care. However, the purpose of output measurement is not to provide estimates of the societal value, so the use of cost weights does not constitute a major drawback in the context of the national accounts.

A challenge: capturing quality change

Health and education services undergo ***quality change:*** existing services are transformed, new services appear and old services disappear. Capturing quality change in volume measures of output is challenging. This handbook draws a distinction between implicit and explicit quality adjustment of volume measures. Implicit quality adjustment happens when products are suitably differentiated or stratified in measurement. In general, the more detailed the stratification, the more similar the processes or activities that are compared over time or in space, and the better the implicit quality adjustment. Suitably differentiated counts of hospital treatments, on the basis of diagnosis related group (DRGs), provide an example of such implicit quality adjustment. Care must be taken, however, to ensure that stratification is reflective of outputs, not inputs.

Implicit quality adjustment may not be always be sufficient however and methods of explicit quality adjustment may be required. For this purpose, it may be necessary to invoke outcomes. Indeed, the notion of 'quality' of health and education services is hard to define without *some* reference to outcomes. The most prominent example in education services is quality adjustment through exam scores. Exam scores are the joint outcome of teaching, student effort, natural ability and the broader socio-economic environment. If those changes in scores that are due to teaching *only* can be identified and measured, estimates of the quality adjustment of education services are obtainable. While not straight forward, methods to identify the effects of teaching on exam scores exist, and the handbook discusses these in some detail. It is important to stress however that it is (quality-adjusted) output, and not outcomes as such, that is the target.

In health services, explicit quality adjustment is also complex. Many health care quality indicators exist, including those developed at the OECD but health care quality is multi-dimensional and there is no universally agreed method to aggregate across different quality dimensions nor is there an agreed approach on how to isolate the effects of the medical care on health outcomes. For these reasons, the handbook discusses but does not put forward proposals for explicit quality adjustment in the case of health care services.

Consistency between temporal and spatial comparisons

Two chapters of the handbook relate to cross-country comparisons of health and education output. Such spatial comparisons require health and education-specific ***Purchasing Power Parities (PPP)***. These currency conversion rates are then applied to health and education expenditures of different countries to obtain international volume comparisons of health and education services. Computing PPPs is by definition a task that has to be carried out at the international level and the respective parts of the handbook describe the projects undertaken by the OECD (in co-operation with Eurostat) in this area. To date, PPPs for education services are available but PPPs for health services are still under development. In both cases, the spatial approach is broadly consistent with the temporal approaches put forward in this handbook.

Aligning statistics with administrative requirements

Many of the datasets crucial to developing volume measures of health and education services are administrative datasets, conceived for purposes other than national accounts. For instance, the primary purpose of disease-based cost information is accountability for costs and transparency of resources going into health provision. Usability for national accounts is a welcome by-product but not always a planned by-product. The challenge for statistical offices is to help create as many planned by-products as possible. This is not always straight forward because statistical requirements may be different from administrative requirements. Attention will have to be given to aligning statistical frameworks and strategies of economic statistics with health, education and welfare statistics in order to provide a common basis for organising administrative data.

…but more work will be needed

The title of this handbook deliberately points to the fact that further work is needed. It provides guidance in many areas and identifies best practice within current data constraints but as experience is gathered and research progresses, it will certainly be necessary to update and improve the present text. In this context, the handbook's ambition is to clarify concepts, provide a reference for existing work in countries and to propose measurement avenues.

INTRODUCTION AND BACKGROUND

Objectives

0.1 This Handbook aims at providing guidance on the measurement of the volume of education and health services. National accounts provide the reference framework but it is hoped that some of the considerations will be of use beyond the national accounts.

0.2 The Handbook is about the measurement of output of the main products of two industries, education and health services, large parts of which are often government-provided. In the pioneering age of national accounts, the fact that government services contributed to GDP was not accepted by all. However, since the 1968 System of National Accounts, non-market production has been fully recognised as contributing to GDP, even if, as there are neither sales nor market prices, the estimation of their production at current prices has to rely on costs and at constant prices using deflated costs.

0.3 This approach created a major obstacle to the measurement of productivity growth in the non market sector and in 1975, Peter Hill developed the principles for volume measures of non-market services, in particular health and education services (Hill 1975). The approach presented in this Handbook is fully compatible with Hill's work. The *UN Manual on Volume Measurement* also authored by Hill (Hill 1979) drew on his earlier work and provided a first official guideline for output measurement of non-market services.

0.4 Despite the widespread recognition that input based measures presented obstacles, a systematic effort at the international level was not undertaken until 2001 when Eurostat set out guidance through its Handbook on Price and Volume Measures in National Accounts for both market and non-market services. The Eurostat Handbook became part of European law, obliging member states to implement its recommendations. Despite the guidance provided in the handbook, it became clear that inconsistent treatment between countries could easily arise. This issue was also recognised in the Atkinson Review (Atkinson 2005) in the United Kingdom. The Atkinson Review both assessed measures of government output in the U.K. and developed general guidance and measurement principles on this matter.

0.5 The present Handbook builds on the above work, with two special added values. The first is that it is the result not only of discussions between experts of national accounts, but also of the experts participating in the specialised OECD networks for education and health. This synergy should ensure the relevance and practicability of the recommendations of this handbook. The second added value is that it simultaneously considers the temporal aspect (analysis for a particular country over time) and spatial aspect (analysis across countries at a particular point in time).

Measurement for different purposes: national accounts and performance indicators

0.6 There are several questions associated with the performance of the health and education sector and, typically, different questions give rise to different data requirements. Häkkinen and Joumard (2007) distinguish the following levels of analysis.

0.7 First, analysis of the efficiency and cost-effectiveness of the health or education system *as a whole* is a topic of policy-relevance. Broadly speaking, this implies measuring those changes in the health status of the population or in the state of knowledge and skills of the population that can be attributed to public spending on health care or education. Volume measures of health and education services in the national accounts would only be of partial usefulness in this analysis: the health or education system comprises more than those economic units that figure under the health or education industry in the national accounts. For example, an anti-smoking campaign may have a positive effect on the health status of the population but would not necessarily figure as output of the health industry in the national accounts. By the same token, the health effects of introducing safety features into cars are not counted as provision of health services by the car industry.

0.8 Most governments are also interested in improving government efficiency through performance measurement. These targets, such as the number of people on waiting lists for treatment, are generally measurable in a quantifiable way and are often complementary with the objective of measuring the output of non-market services. For example, when performance indicators are expressed in units that are correlated with the measure of consumer satisfaction (delay for hospital surgery, success in exams, etc.) they offer the potential to be used for quality adjustment in the national accounts. However, some care is needed to ensure that the explicit contribution of the non-market sector is captured. For example some performance indicators measure directly the health or educational status of the population overall; changes in which may not be entirely attributable to government services.

0.9 Second, the analysis of efficiency and cost-effectiveness at the level of *individual diseases or individual levels of education* constitutes another area of interest. This analysis implies measuring those changes in the health or education status that are attributable to health care or educational services, wherever performed in the health care or education system. Volume output measures of the health and education industry are better suited for this purpose than for system-wide analyses. Problems exist nonetheless because there is no straightforward way to track treatment of diseases or educational activity across institutions. In particular, movements between in-patient and out-patient care in the health sector poses a challenge for the use of national accounts information.

0.10 Third, there is the analysis of efficiency at the level of *individual diseases or individual levels of education for a given institution*. This constitutes the most limited perspective; particularly in the health sector where it is difficult to differentiate between in-patient and out-patient services. At the same time, the institutional unit or establishment is the building block in the national accounts and consequently the starting point for the search for statistical information. Moving beyond individual institutions towards tracking treatment of diseases or educational services across institutions is possible but challenging as will be explained further in this *Handbook*.

Structure of the Handbook

0.11 In the measurement of services, and more so non-market services, a clear presentation of concepts and an unambiguous terminology are required. Chapter 1 of the Handbook starts out, therefore, with terminology and concepts. Chapters 2 and 3 deal with education services, from a temporal and from a spatial perspective. Chapters 4 and 5 tackle health services.

CHAPTER 1. TERMINOLOGY AND CONCEPTS

1.1 Production and consumption of education and health at current prices

1.1 ***Institutional units.*** An important aim of the National Accounts is to value production, the transformation of inputs, such as labour or capital, into outputs in the form of goods and services. The task of measuring production begins with identifying the units which produce the goods or services and, particularly for education and health services, where a significant proportion of output is non-market, those units that are non-market producers.

1.2 Market output is output that is sold at prices that are economically significant. Thus, for market services of education and health, the value of output in current prices can be measured by the value of sales of these services.

1.3 However, education and health are the most common examples of services provided by government free of charge or at prices which are not economically significant and thus constitute non-market output. A price which is not economically significant is deliberately fixed well below the equilibrium price that would clear the market. The SNA defines it as a price which has little or no influence over how much the producer is willing to supply and which has only a marginal influence on the quantities demanded.

1.4 There are differences in country practices to identify the economic significance of prices. For instance, the European System of Accounts (ESA 1995) considers, for practical reasons that a price is not economically significant if it covers less than half of the costs of producing the service. Whatever the exact rule, valuation of output is based on adding the costs incurred in production; namely the sum of:

- Intermediate consumption (the goods and services used up in producing the service)
- Compensation of employees (costs of teachers, doctors, nurses, etc…)
- Consumption of fixed capital[1] (depreciation of school and hospital buildings, of medical equipment)
- Other taxes, less subsidies, on production.

1.5 ***Industries and products***. An economy's supply side can be described by the composition of its products or of its industries. These two categories are linked as the products produced determine the industry allocation of producers and any calculation of volume and price for an industry needs a break down by product. Producers are allocated to industries according to a classification system of industries twinned with a classification system of products. The classification systems most commonly in use are based on the International Standard Industrial Classification (ISIC) linked with the Central Product

---

[1] The cost of capital comprises more than depreciation. In particular, there are financing costs and there may be costs from revaluation of assets. However, by convention, the national accounts only recognise depreciation as the cost element for capital held by non-market producers. For a discussion see OECD (2009) and Atkinson (2005).

Classification (CPC). The classes of ISIC rev 4[2] which are within the areas covered by this Handbook are Education (Division 85) and Human Health Activities (Division 86), an extract of which is shown in Table 1. This Handbook focuses on formal education (classes 851-853) and human health (division 86), but its guidance may be relevant for other activities, such as "other education" or "residential care".

Table 1.1. Health and education in the industry classification industries (extract)

Education (Division 85)	
851	Pre-primary and primary education
852	Secondary education
8521	General secondary
8522	Technical and vocational secondary
853	Higher secondary
854	Other education
855	Educational support activities
Human Health (Division 86)	
861	Hospital services
862	Medical and dental practices
863	Other human health
Residential Care (Division 87)	
871	Residential nursing care facilities
872	Residential care activities for mental retardation, mental health and substance abuse
873	Residential care activities for the elderly and disabled
879	Other residential care activities

Source: International Standard Industrial Classification, rev 4.

1.6 ***Final expenditure, supply and use tables***. In the traditional framework for final expenditures, the national accounts distinguish between the types of consumers: final consumption expenditures of households; final consumption expenditures of non profit institutions serving households (NPISH); and final consumption expenditures of general government. The allocation of expenditures to these groups is based on *who pays* or more precisely, who incurs the direct expenditure. If households (for example pupils and patients) pay, the expenditure is classified as final expenditures of *households*, if it is the government (which includes social security) the expenditure is classified as final expenditure of *government,* which includes any reimbursements made by government to households for services they procured. This classification is not well suited to analyse the *consumption,* as opposed to the *expenditure*, of education and health services by households, i.e., when the question is 'who consumes?' rather than 'who pays?'. For example, free or quasi-free education and health services are paid for by government on behalf of households, but they are consumed by households. The national accounts framework includes therefore another aggregate, called *actual individual consumption*, which is the sum of the expenditures made directly by households plus those made by government on behalf of households. Actual consumption is the aggregate that will often be used in this Handbook when the "consumption" of education and health services is discussed.

1.7 There are also classifications of household consumption and of government expenditures. The flows of goods and services between the supply and the demand sides of an economy are captured by supply and use tables. They show, for a product or a group of products, the amount supplied by producers (whether non market or market) and the amounts consumed by households, used for investment or

2 See http://unstats.un.org/unsd/cr/registry/isic-4.asp

exported. The supply-use tables in the national accounts are most relevant for the analysis of the output and, at the same time, the consumption of education and health services.

1.2 Production and consumption of education and health services in volume

1.2.1 Introduction

1.8 Although much effort is spent on measuring the value of GDP at current prices, an often more important objective of the National Accounts is to derive a measure of the growth of GDP and its components *in volume*. Growth in volume controls for changes in the price level between two periods. International comparisons of GDP should also be made excluding differences in the price level between countries. Only elements of the National Accounts that can be disaggregated in terms of prices and volumes are useful in analysing economic growth, productivity and inflation. It is indeed the main objective of this Handbook to deliver recommendations on the measurement of the *volume* of output of health and education services. For complex services such as education and health, this is a difficult task.

1.9 Volume is associated with quantities. To express a flow in volume terms, each of the goods and services which are the counterparts of money spent must, in principle, be identified. This is because quantities are additive only for a single homogeneous product: if a hospital carries out 100 (expensive) knee replacement operations and 300 (inexpensive) varicose vein treatments, it is not informative in economic terms to add these together and say that 400 treatments were carried out. Clearly there is a vast number of different goods and services of varying specifications so our search for items to quantify has to be limited in practice and this imposes limitations on how the results are interpreted. But the lack of a detailed specification for each and every item produced or consumed is not a sufficient reason to invalidate this approach: it just has to be applied with caution and a good understanding of what is being measured.

1.10 National accounts are about constructing macroeconomic aggregates. Inevitably, the question therefore arises of how to add together the quantities of the very detailed homogeneous products that were alluded to in the previous paragraph. To use the same example, how are the 100 knee replacement operations and the 300 varicose treatments to be added in a meaningful way? This question will arise for all the thousands of goods and services that populate the global concept of education and health services. As will be seen, the answer traditionally draws on the knowledge that the relative prices of the different goods and services bought and sold reflect both their relative utilities to purchasers and their relative costs for the producers. However, is this relevant in particular for non market services where, by definition, there is no market price that reflects the interaction of consumers and producers? This is discussed next.

1.2.2 Inputs, Output and Utility

1.11 The conceptual discussion starts with a simple market model of producers and consumers. On one side of the market, there are producers who supply goods or services that are the result of a production process. In the production process, labour, capital and intermediate inputs are combined with a certain technology to produce outputs, the products that are typically destined for transactions. For the moment, outputs are taken as well-defined and transactions of these outputs on the market are taken as observable. Note the link of outputs to transactions. In the simple market case, these will be market transactions where money is exchanged for a product. In the non-market case, the form of transaction will be different but there will always be a transaction or a transferral of products. This is of importance because it can help determine the location of the production boundary and the conceptual place to measure output. This was already clarified by Hill (1975): "[…] output is measured in terms of transactions between producers and users, and the production boundary is drawn at the point at which the producer unit actually sells or otherwise disposes of his output to another economic unit". When there is competition on the market, producers will supply outputs to the point where the price that they achieve on the market equals the marginal cost of producing an extra unit of this output.

1.12 On the demand side, consumers purchase the goods and services supplied. Standard economic theory attributes a utility function to consumers where *utility* depends on the quantity of goods and services consumed. The utility function indicates how the consumer appreciates (in unobserved 'utils') the quantity of products purchased. Utility is difficult to measure because it is a subjective concept and because it is not directly observable. However, for the purpose at hand, the notion of utility is useful in at least two ways.

- First, the utility perspective has implications for the statistician who has to classify goods and services into categories when measuring price and volume developments: price and volume indices, even at the most disaggregated level, nearly always reflect several individual goods or services ('items') that are grouped together as 'a product'. The grouping of items should be performed in such a way that each one satisfies the same or similar consumer needs: this may or may not coincide with production processes. For example, in health services, it may be better to group treatments by type of disease ('treatment of heart attacks') rather than by type of treatment ('medication', or 'operation'). The advantage of this approach is that it compares individual items that are substitutable for each other from the consumer perspective and this has implications for the resulting measure of output, more of which below.

- Second, the utility perspective helps us to model consumer behaviour and to conceptualise price and volume indices. One example of such a concept is the *cost of living index,* described as the extra expenditure that a consumer – in the presence of price changes - has to incur from one period to the next if s/he wants to keep utility constant. Not every consumer price index is a cost of living index but if a price index is designed to reflect changes in the cost of living this will normally have implications for the measurement of health and education price and volume indices.

1.13 In the simple market situation taken here, a price index can be used to deflate the value of products that has been transacted on the market, and this yields a volume measure of the goods or services transacted. Instead of deflating values with a price index, one could also set up an index of volumes or quantities produced and consumed directly. Price and volume indices of the Laspeyres[3] type are shown below where P_i^t stands for the price of product i in period t and Y_i^t for its quantity:

$$(1) \qquad P^{L,t} = \sum^N \left(\frac{P_i^{t-1} Y_i^{t-1}}{\sum^N P_i^{t-1} Y_i^{t-1}} \right) \frac{P_i^t}{P_i^{t-1}}$$

$$(2) \qquad Y^{L,t} = \sum^N \left(\frac{P_i^{t-1} Y_i^{t-1}}{\sum^N P_i^{t-1} Y_i^{t-1}} \right) \frac{Y_i^t}{Y_i^{t-1}}$$

1.14 Three things are worth noting about these indices.

- First, prices or quantities of the products are weighted with expenditure shares and these expenditure shares – measured by market observation – reflect the joint, equilibrium valuation of

3 The Laspeyres-type formula is shown here because it is the most commonly used index number formula. For a discussion of index number formulae for consumer and producer prices, see ILO et al. (2004a, 2004b).

each product by consumers and producers. Thus, market prices and quantities reveal the interaction of consumer preferences and producer costs.

- Second, *in the simple model at hand*, changes in quantities Y_i^t/Y_i^{t-1} and in prices P_i^t/P_i^{t-1} are simply measured by comparing them between periods – implying that the units of measurement for Y_i^t are the same as the units of measurement for Y_i^{t-1} and that the set of products is stable – product i has to exist in both periods to be compared. A volume index of goods could thus be constructed as a weighted average of the number of goods transacted and a volume index of services could be constructed as the weighted average of the number of actions or activities transacted in the two periods. The same is true for a price index, which could be constructed as a weighted average of the price changes between two periods. No further reference is needed to notions of utility and this corresponds to the general guidance given in the System of National Accounts. The picture changes when there is quality change and when new products appear. Matters are further complicated when it comes to the specific areas of health and education services because prices may not represent an equilibrium valuation and in some cases, no price or an economically insignificant price is charged.

- Third, the simple presentation here also makes the implicit assumption that there is exactly one measured unit of quantity that constitutes Y_i or one measured price that constitutes P_i. It was already mentioned earlier that in practice, this is rarely the case. The P_i's are un-weighted averages of individual items which constitute an elementary price index. Similarly, the Y_i's for a volume index are actually un-weighted averages of quantities of individual products. How individual products ('items') are grouped is a question that has to be answered with respect to the purpose of the price or volume index. Above, it was mentioned that a useful criterion for grouping individual items is that they potentially satisfy the same or similar consumer needs or that they are substitutes from a consumer perspective. Conversely, if different items are not interchangeable from a consumer perspective, they should be treated as different products. In the presence of quality change or new and disappearing items, the question of grouping items becomes important, more of which below.

1.2.3 Quality change, new and disappearing products

1.15 An unrealistic assumption in the model above is the set of unchanged products between two periods. In reality, the quality of products changes over time, certain products disappear from the market and new products emerge. These changes constitute not only a major practical challenge for statisticians they also have consequences for theoretical considerations about output and utility. The distinction between new products and quality change[4] will be ignored here but a few general points about quality adjustment[5] of prices or quantities will be noted.

1.16 One technique to deal with quality change in products is to group them such that only products of the same specification are compared over time or in space. Such grouping or matching ensures that only prices or quantities of products of the same or very similar quality are compared. The idea is that products of different quality are treated as different products. Examples for such grouping in education are establishments that provide different services in addition to education, such as boarding schools as opposed to day-time schools or hospitals with different levels of non-medical services. Note, however, that grouping also relies on an important assumption: to show a price or quantity movement that is representative of a product group, the price or quantity movement of those products that *are* matched has

4 For a discussion see for example ILO et al. (2004).

5 For an in-depth treatment of quality adjustment in price measurement see Triplett (2006).

to be a good indicator of the price or quantity movement of those products that are *not* matched – in particular, products that are newly entering the market. Also, all other price or quantity changes that arise outside of the sample of matched products are ignored.

1.17 If matching is insufficient, other explicit, techniques have to be invoked to account for quality change. In general, the quality of a product can be expressed by the quantity of its characteristics. Quality change can then be captured by the change in characteristics. Similarly, price changes in products can be attributed to pure price changes and to those price changes that reflect changes in product characteristics. This is the approach followed by hedonic price indices[6] that are now well established among statistical agencies.

1.18 Quality adjustments require the identification of a set of characteristics such as the speed, engine size or equipment for a car or the processor speed for a computer. Berndt et al. (2001) use patient characteristics, information on different types of depression, variables on medication and the like to estimate a hedonic price model for the treatment of depression; the idea being to isolate those price changes that are due to changes in characteristics from those price changes that constitute 'inflation'. An important result of the hedonic model is that it allows the identification of characteristics and provides a market valuation of each one[7]. Market valuation, in turn, is a convenient way of aggregating across characteristics because everything is expressed in a single monetary unit.

1.19 While hedonic regression techniques can help to value characteristics, there are situations when there are no market prices or when there are good reasons to believe that market prices do not reflect consumer preferences or costs for producers. The first case may arise when services are provided by non-market producers, i.e., at prices which do not cover costs of production. The second case may arise when consumers and producers interact only indirectly. For example, individuals take out health insurance. Then, the price of a medical intervention is not a signal that is directly relevant to consumers although it will be a relevant signal to insurers and to health providers. The implication is that such prices may not be useful in revealing consumer preferences.

6 See Triplett (2006) for a comprehensive discussion.

7 Rosen (1974) demonstrated that in general, those characteristics of a product will show up in the function that is valued by consumers *and* that has cost implications for producers.

Box 1. Narrow specifications of products and quality change

A straight forward technique of dealing with quality change in a price or in a volume index is to match models, i.e., to compare only prices or quantities of products that are tightly specified. In other words, products are treated as different products whenever their characteristics are different. The more specific the characteristics of a particular product, the less likely it is that a modification of the product goes unnoticed and that a change in quality is not recognised as such. Such implicit quality-adjustment is well adapted when the set of products observed is stable and when it is representative for the universe of products. It may, however, be insufficient, when products change, when there are substitution processes between new and old products and when there are no markets or when existing markets operate imperfectly. This is best illustrated by way of an example. A quantity index is used here but the same points could also be made by way of a price index that is subsequently used to deflate values.

Suppose there are two treatments for a disease, traditional surgery and laser treatment, and assume that laser treatment is introduced in period 1. In addition, as may well be the case, the unit cost of laser treatment is lower than the unit cost of traditional surgery. The total number of interventions in each period remains the same.

	Traditional surgery			Laser surgery		
	Period 0	Period 1	Period 2	Period 0	Period 1	Period 2
Unit cost	100	100	100	-	90	90
Number of interventions	50	40	5	0	10	45
Total cost	5000	4000	500	0	900	4050

Now consider a (simplified) *matched-model approach* towards calculating a volume change from period 0 to period 1. In the simplest case, the volume index is given by the quantity changes in the two treatments, each weighted by the cost share it occupies in period 0. As laser surgery does not yet exist in period 0, it receives a zero weight so the volume index of treatments is simply the change in the number of traditional surgery interventions, or (40/50-1)=-20%. Between periods 1 and 2, the corresponding volume index equals $[s_T(5/40)+s_L(45/10)]-1=-7.1\%$ where $s_T=82\%$ and $s_L=18\%$ are the period 1 cost shares of the traditional and laser treatments respectively. This approach treats the two treatments as different products and the sharp drop in the total volume index in period 1 reflects the 'new goods' problem that arises when new products enter the sample that cannot be compared with quantities in the base period. The implicit assumption in this model is that consumer valuation of the two products is captured by the relative unit costs, so if laser surgery is cheaper than traditional surgery, the method implicitly quality-adjusts *downward* the quantity of laser surgery when combined with traditional surgery. In a perfect market, the price of the traditional treatment would see an instantaneous downward adjustment, bringing consumer valuation of the two processes in line.

A different result arises when it is considered that the two treatments are perfect substitutes, i.e., that they are in fact the same product. In this case, no cost weighting is applied between the two treatments – and the number of treatments is simply added up. As there are 50 interventions in every period, the result is a volume index that shows zero growth and a declining price index, reflecting the drop in average unit costs of treatment.

The previous method is justified if consumers are indifferent to the two treatments. If this is not the case, and they prefer laser over traditional surgery because the former is less intrusive or requires fewer days of recovery, an explicit quality-adjustment is needed. Such an adjustment can be applied to the quantity measures, either by scaling up the quantity of laser treatments or by scaling down the quantity of traditional treatments. Whichever way this is done, the implication is always that one treatment is expressed in equivalents of the other, and the ratio should in some way reflect consumer preferences. Alternatively, prices or unit costs could be rescaled before constructing a price index. Suppose the adjustment factor is 1.1 – each laser treatment is the equivalent of 1.1 traditional treatments. Then, expressed in 'traditional surgery-equivalents', the number of treatments is 50 in period 0, 40+10*1.1=51 in period 1 and 5+45*1.1=54.5 in period 2. The resulting volume index is +2% in period 1 and +6.9% in period 2. Obviously, the difficulty lies in determining the adjustment factor which should (i) reflect consumer preferences; and (ii) be uni-dimensional. Much of the present *Handbook* is actually devoted to the identification and measurement of such adjustment factors.

(Note that the above example is simplified to make the central point about substitution. The result of -20% is actually only a lower bound to the Laspeyres index because strictly speaking the volume index is undefined.)

1.20 What remains true, however, independently of whether or not goods or services are transacted in a functioning market, is that when products undergo quality change, when some products disappear and others emerge, it is always necessary to identify characteristics of products. Note that typically, there is more than one quality characteristic to a product. If there are no market signals to identify relevant product characteristics, the statistician has to rely on expert advice to identify the characteristics that consumers and/or producers value. Health and education experts have indeed produced sets of characteristics that they consider as quality attributes of health care or of education services. But even a well-defined and well-measured set of characteristics leaves the problem of aggregation across the quantities of characteristics which are, by their very nature, expressed in very different physical units. If there is a need to produce a single number for quality adjustment, and if there is no market information to value characteristics, measurement has to proceed differently. Enter the notion of *outcome*.

1.21 Consumers attach utility to a good or to a service because it affects *outcome*, i.e., a particular state that they value and which can be measured. One could also say that outcome is an intermediate step between consumption and utility and this is indeed the way it has been treated in the literature. For example, Berndt et al. (2001) distinguish between medical care ('output' in our terminology), the state of health ('outcome' in our terminology) and utility. They envisage a relationship whereby utility depends, among other variables, on the state of health and where the state of health is itself dependent on health care services, on the environment, lifestyle etc.). Thus, a health activity with a higher composite quality than another health care activity could be identified as such if it contributes more to health outcome than the alternative activity[8].

1.22 Thus, one possibility to deal with the aggregation problem is to subsume several characteristics into a single indicator that reflects the *contribution of the product to outcome*. For example, in the case of price indices for health care, Triplett (1999) suggests quality-adjusted life years (QALYs) as a single dimensional measure that could be used for the quality-adjustment of different treatments within a product group. Cutler et al. (1998) derive a price index of heart attack treatments using QALYs. Chandler (2009) shows how an outcome production function can be used to estimate output as the marginal contribution to outcome. De Haan (2009) also reasons that output of the education sector should be measured as the marginal contribution of education services to outcome, and discusses methodology. In all these cases, a central point is to derive a single indicator that serves as a reasonable summary of a true, multi-dimensional set of quality characteristics valued by consumers when purchasing health or education services. Careful judgement needs to be applied in the choice of such a measure. In particular,

- it should not be affected by any other factors that influence consumer outcome (e.g., socio-economic background of students or lifestyle of patients) and

- it should reflect as closely as possible the normal, or average or expected effect of the activity on the state of health or the state of knowledge of the consumer. 'Normal', 'average' or 'expected' has been added here to signal that to the extent possible, measures of service production should not be influenced by the individual capacity of the consumer to make use of these services. For example, the same teaching activity performed on a different group of students, should be measured as the same quantity of teaching services. Or the same treatment, applied to two different persons with the same disease should be measured as equal quantities of medical services.

8 Things are more complicated in concept and in practice. First, as Berndt et al. (2001) point out, there is an issue of lags: the state of health may be affected by medical care and by other factors with a lag so that utility derived from the state of health occurs at a different date from when medical services are provided. Second, there may also be a trade-off between immediate utility derived from consumption (say a fatty diet) and long-term disutility from reduced health status. This complicates formalisation of consumer behaviour but is secondary to the issue at hand, namely the measurement of health services.

1.23 Note that the last principle – measuring output without regard to the individual capacity of the consumer, is not straight forward to apply. In principle, it should not matter for the volume measure of services *who* is consuming a service just as the volume measure of car production should be unaffected by whether a car is purchased by an old or by a young person. However, when the nature of the service changes due to certain consumer characteristics, the situation may be different. For example, an elderly patient suffering from the same disease as a young patient may need more care due to longer time to recover. This may result in higher expenditures for the group of older patients. Similarly, providing education services for special types of pupils might be costlier than for the average pupil. In such cases, when grouping together services with different consumer characteristics, a higher volume rather than a higher price should be recorded for those characteristics which incur higher costs, because the nature of the service is different. There are thus cases of different products with the implication that volume measurement should be stratified by type of consumer[9].

1.24 To sum up, it is necessary to qualify the earlier statement that to measure the quantity or the price change of a single product (P_L^t, or Y_L^t), one does not have to recur to considerations of outcome. This gives rise to the following conclusions.

1.25 A first step towards capturing quality change is the correct stratification, i.e., the comparison of products with the same or at least similar characteristics. This way, quality changes are automatically controlled for. However, matching of services has its limits when comparable products do not exist in comparison periods. Care must be applied in defining products to be compared so as to capture effects of substitution (see Box 1) without, however, treating goods or services as substitutes that are in fact different products (such as health treatments specific to particular parts of the population).

1.26 Second, for the process of quality adjustment of the quantities or prices of output it may be necessary to invoke outcomes, because characteristics that matter for consumers have to be identified for quality adjustment. Thus, it may be necessary to look beyond the strict national accounts production boundary if one wants to deal with quality change in health and education services. This does not mean that the production boundary is shifted but it indicates that the quality of services cannot be identified without considering indicators falling outside the production boundary.

1.27 In general, quality is multi-dimensional. Some of the multi-dimensionality can be captured by stratification. For certain market health or education services, it may be possible to construct hedonic price indices that permit combining quality attributes into a single monetary measure. In many instances, this will not be possible and explicit quality adjustment has to be based on a single quality characteristic, which is directly linked to the impact on the "outcome". The choice and measurement of this characteristic is key in the process and needs careful consideration.

1.28 With the above remarks on quality adjustment in mind, volume measures of output can generally be defined as the quality-adjusted counts of processes. Processes should be classified by consumer-relevant categories such as the completed treatment for a particular disease or the level of education provided. These definitions will be elaborated on more in the relevant chapters.

9 There is nothing extraordinary in the assumption that the "same" service is different for different consumers. For example, products intended for children are often smaller than those for adults (clothing, restaurant meals etc.) and are treated as separate products in price or production statistics. A similar quantitative difference in products may concern also services.

Box 2. The meanings of 'outcome'

Outcome has been used in different ways in the relevant literature on health services. Two usages are common:

In the health care literature, 'outcome' is typically defined as the resulting change in health status that is directly attributable to the health care received. Triplett (2001) indicates this usage in the cost-effectiveness literature and quotes Gold et al. (1996) who define a health outcome as the end result of a medical intervention, or the change in health status associated with the intervention over some evaluation period or over the patient's lifetime. Employed in this sense, some authors suggest that the 'output' of the health care industry be measured by 'outcome'.

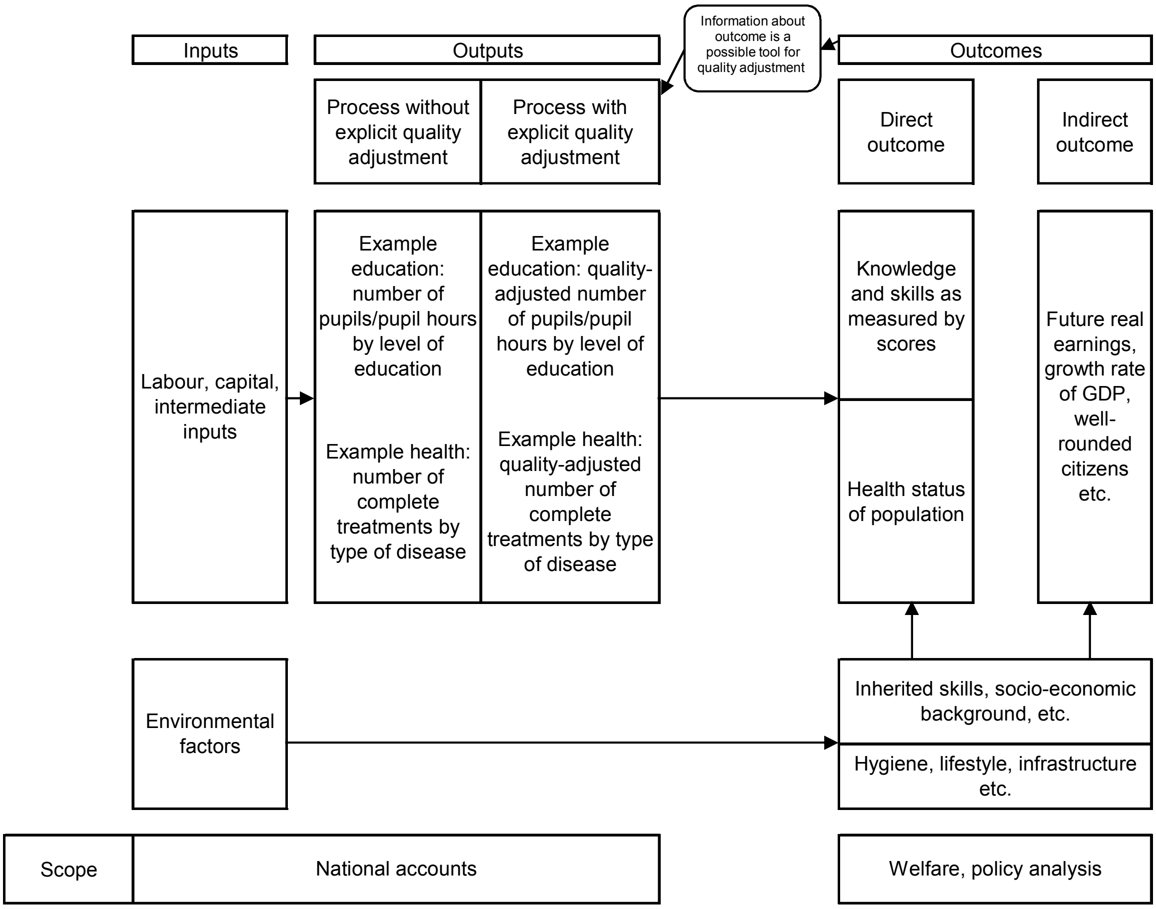
Among national accountants, 'outcome' is typically used to describe a state that consumers value, for example the health status without necessarily relating the change in this state to the medical intervention. For example, Eurostat (2001) gives as examples of "outcome indicators" the level of education of the population, life expectancy, or the level of crime. Atkinson (2005) has the same usage of the word. Understood in this sense, outcome in itself cannot be a useful way to measure output or the effectiveness of the health or education system. In terms of national accounts semantics, the 'marginal contribution of the health care industry to outcome' is the equivalent to the notion of 'outcome' as used in the health care literature.

As long as a particular definition is used consistently, the substance of the argument is of course unaffected and the only question is the usefulness of one definition or the other. As the present handbook follows in the line of Eurostat (2001) and the Atkinson Review (2005), it also employs the term 'outcome' in the sense of the national accounts literature.

1.29 In the context of measuring health and education output, it is useful to refine the broad distinction between inputs, outputs and outcomes in two ways. First, outputs are broken down into two components: activities or processes and the quality adjustment applied to them. Processes are observable actions by which health and education services are delivered although their characteristics may change over time. For education, a typical process measure is the number of pupils or the number of pupil hours taught in a particular grade. For health, a typical process measure is the number of treatments of a particular disease such as hip replacements.

1.30 Second, outcomes can be broken down into direct and indirect outcomes, the distinction being that direct outcomes are closer to the act of service provision than indirect outcomes although, in line with the discussion above, neither direct nor indirect outcomes are measures of services. For example, in the case of education, a direct outcome is the state of knowledge of a population of pupils, estimated by scores or degrees. The indirect outcomes associated with education are employment possibilities and enhanced real earnings due to better education, or GDP growth as a consequence of enhanced human capital. Indirect outcomes associated with health services are fewer working days lost due to diseases, or individual well-being. These distinctions between activities, quality adjustments and direct and indirect outcomes are shown in Figure 1.1 below. The figure also depicts the scope of national accounts measures which are defined via the production boundary. However, as explained above, information about outcomes, in particular about the contribution of health and education services to health and education outcomes, can provide a tool for the explicit quality adjustment of processes or activities.

Figure 1.1. Inputs, outputs and outcomes



1.2.4 Cost and value weights

1.31 The considerations above already allude to the fact that in a non-market context, prices for goods or services are not in general observable. It could be argued that this may also be the case for market services where a price mechanism is observable but not necessarily a price per unit of output. For example, if health services are defined as completed cycles of treatment, the price for a particular completed treatment may not be directly observable, because the observable pricing mechanism may be geared to individual activities (physician services, laboratory services etc.) rather than to a complete treatment. There is also the issue of insurance: prices that are observed in the market of health services may not be indicative of consumers' willingness to pay if these are covered by health insurance. This leads us to conclude that differences in measurement between 'market' and 'non-market' provision of health and education services may be less pronounced than is sometimes assumed.

1.32 Whatever the precise difference between market and non-market set-ups in the health and education sectors, it is evident that volume or price indices of the kind shown in equations (1) and (2) necessitate a weighting system and, for many products that are transacted on a market, observed revenue or expenditure shares constitute the obvious choice for such weights. However, in the absence of market prices or in the absence of prices that constitute signals to market participants, the relative valuation of a product by the consumer may differ from the relative costs of producing it. There are several ways of dealing with this situation:

- *Assume* that, on average, the relative valuation by the consumer equals the relative cost incurred by producers. For government producers, the argument is that in a democracy, and via the electoral process, consumers exert some influence over the production decision of governments so that the provision of non-market services is socially optimal, at least over longer periods. In this case, the equality between relative costs and relative utility or willingness to pay holds approximately and all that one needs to do is look for the empirical measurement of one of the elements. Nearly always, costs will be easier to come by and so constitute the first choice from a practical perspective.

- *Impute* some value for the *relative valuation by the consumer* to generate aggregation weights. The imputation of consumers' valuations of certain medical or educational services implies embracing a fully developed consumer or welfare perspective in the estimation of the volumes of health and education services. This may be controversial from a national accounts perspective if it implies that the total value of health or education services is different from the total expenditure for these services, or if the construction of weights entails delicate questions such as putting a monetary value on human life[10].

1.33 Which approach is preferable? For several reasons, it would seem preferable to stick with the supply side and use weights that are based on costs rather than on consumer valuation. This is broadly consistent with the theory of price indices (Fisher and Shell 1972) where output (price) measures are based on revenue weights (that equal marginal costs) and where input (price) measures are based on consumer valuation. On the other hand, there is also a consumption component to government output (incidentally, 'government output' and 'government consumption' are often used interchangeably) which suggests looking for weights that reflect consumer valuation.

1.34 Theory favours cost weights for output measures and there are also pragmatic reasons for this choice. Cost shares have the significant advantage of greater measurability than utility-based weights. In addition, using the cost shares as an imputation for the revenue shares of non-market output implies that the equality of total costs and total revenues at current prices is kept during every accounting period. This is in line with a principle of national accounts for the measurement of non-market output that stipulates that the value of non-market production equals the value of its inputs. If a different valuation of outputs were used, the total value of non-market production would not in general equal total costs. While this may not be a problem as such – in particular when non-market production is viewed from a pure welfare perspective – it poses a practical issue of dealing with an additional item in the national accounts ('social surplus' or 'social loss' arising from non-market production).

1.35 Another argument rests on the fact that if statisticians attempt to attach full-scale user values to changes in product characteristics for health and education services, this may introduce a bias with regard to the treatment of other products in the accounts. For example, the reduction in mortality and the associated gain in lifetime income that arise from the introduction of airbags in cars are not normally taken into account in the process of quality-adjusting car prices.

1.36 In conclusion, when there are no or inadequate market prices, there is no guarantee that relative consumer valuation and relative costs of a product coincide. Value weights and cost weights will therefore yield different results. However, in the absence of strong conceptual reasons against cost weights and in

10 Using value weights for a price or quantity index does not necessarily imply putting a monetary value on a year of life. This may be required for cost-effectiveness studies (see Triplett 1999). However, basing weights on consumer valuation implies at a minimum, that indicators of outcome, for example QALYs in the case of health care, are compared across health service products.

the presence of many practical reasons in favour, cost weights emerge as the best way towards implementation.

1.2.5 Times series and cross-country analysis

1.37 The exposition has so far centred on measuring the volume output of education and health in a time series setting: this measures output growth between two time periods while holding the unit covered constant. In other words, it measures output growth in one country or part of a country *over time*. The principles set out earlier in this chapter can also be used to measure output across countries while holding time constant, that is carry out a "cross-country" comparisons. The aim should be that two countries producing the same number of identical items of a given quality would be measured as having the equivalent volume of output. The similarity with the method used for time series should be quite obvious: a time series comparison within the same country can easily be seen as comparing two countries.. In other words, the method used to compare the UK in 1995 and the UK in 2005 should be consistent with the method used to compare the UK in 2005 with Turkey in 2005. This similarity will be heavily used in this handbook: the quantity and quality indicators that are recommended for use in a time series analysis will be, at least conceptually, the same as the quantity and quality indicators to be used in cross-country comparisons, and vice-versa.

1.38 National accounts generally calculate volume time-series by deflating current price aggregates by adequate temporal price indices. Similarly national accounts use appropriate spatial price indices ("Purchasing Power Parities") to deflate current price aggregates (typically in different currencies) in order to compare the volume of output (or consumption) between countries. This handbook generally discusses quantity and quality indicators for health and education: number of treatments, QALYs, number of pupils, scores at exams. This may appear to be quite different from PPPs which, as explained, are price deflators, thus weighted price ratios. However, this difference is only presentational, because, as explained in Box 3, volume and unit cost or price indices are reverse sides of the same coin. Indeed, comparing directly the number of treatments between two countries is the same as deflating the cost associated with the treatments in the two countries by a PPP equal to the ratio of the unit cost of the given treatment in each country.

1.39 Thus the objective of this Handbook can equivalently be seen as constructing a set of PPPs for health and education *that correspond to an output based concept of volume*. Indeed, today, the PPPs that are used for the deflation of health and education non market services are input-based PPPs. They consist of *input cost ratios,* essentially labour cost ratios: in one, the costs of one hour of a doctor, nurse or teacher in each country are compared. The new PPPs that this handbook proposes are unit cost ratios for *outputs*: they are the ratio of the unit cost of a given diagnosis related group (DRG) treatment or of a given year of education, where the quality of the treatment or of the year of education is controlled for.

1.2.6 Ensuring consistency between market and non-market measures

1.40 Having dealt with some of the conceptual issues in the previous section, the following discussion takes a look at measurement issues. More details on measurement of health and education services can be found in the relevant chapters of this Handbook.

1.41 Much of the discussion in this Handbook is about the measurement of *non market services.* But this should not distract from the principle that the measure of the volume of education and health consumed by pupils and patients should not be affected *at all* by the status, whether market or non market, of the provider of the service. In other words, the measured volume of *non market services* should be the same as the one for measurement of the volume of *market services*, and vice-versa, as long as the services are the same. This has not been the practice in the past: there has been a tendency to create separate indices for market and non-market production. One objective of this Handbook is to encourage the compilation of

1.42 consistent measures of health and education output, whether these services are provided by market or by non-market producers. Hill (1975) formulated this idea as follows:

> "It is proposed as a matter of principle that the basic methodology used to measure changes in the volume of real output should always be the same irrespective of whether a service is provided on a market or on a non-market basis. This is not to say that the actual numerical measures would not be affected by whether the service is market or non-market, because different weighting systems would be involved, but at least the methods of measurement should be conceptually similar" (page 19).

1.43 To achieve such convergence, an obvious first approach would be to simply apply the deflation method used for the market sector to the non market sector. Indeed, the existence of expenditure by households implies that there should be a Consumer Price Index calculated for these flows, and thus it seems there already exists an experience in calculating a volume/price split and the question arises whether this experience cannot be directly applied to the non market sector.

Box 3. Price indices, unit cost indices and volume indices

When a volume change has to be measured, this can in principle be achieved in two ways, by dividing the change of a value by a price index (deflation) or by directly constructing a volume index. A volume index is a weighted average of quantity changes of individual (homogeneous) products where the amount spent on each service provides the weight. In practice, statisticians lean towards deflation methods because the sampling for a price index is easier to undertake than for a volume index (prices tend to follow more similar trends than volumes) and because it is easier to deal with exiting and entering products

In a non-market context, things are more complicated. Price indices may not exist or may not be meaningful for deflation when there are no economically significant prices. Also, value measures (of output) are typically the sum of costs. In such a case, direct volume indices constitute a valid option and have been used in statistical practice. Alternatively, unit cost (costs per unit of output) can be constructed for purposes of deflation. In this case, unit costs – the total costs of a particular product divided by the number of products of the same type – become a substitute for prices. Note that applying a unit cost index to a change in the value of non-market production (which equals total costs) is equivalent to constructing a volume index directly. The present handbook uses the expressions 'unit cost index' and 'quasi price index' interchangeably.

1.44 While possible in some cases, the extension from the market sector to the non market sector in general is more complicated. A problem occurs when the services provided by the market sector are different from those provided by the non market sector. For example, in some countries, the hospital market sector may be mainly dealing with small or standard interventions, for which it is easier to calculate a volume measure, while public hospital services also cover more complicated interventions, for which volume calculations are also more difficult. Thus, the market sector may not be representative for the entire sector.

1.45 Another reason is that even for *market operators,* the principles for measuring prices and volumes of health and education services are not very well well-established. The existence of a CPI or PPI for market education or health services does not automatically imply that this CPI or PPI is a perfect price index. As explained earlier, there is as much difficulty in taking account of quality change in a price index as in a volume index. Price statisticians have long recognised these difficulties.

1.46 The United States are a case in point. A large part of expenditures on health in the US is directly attributed to household expenditure and provided by "market producers". These expenditures are therefore covered in the CPI. The volume of health services in the national accounts is then obtained by applying a mixture of relevant CPI and PPI deflators to health expenditure. The choice is based on which deflator provides the best coverage. However, the extensive literature on the difficulty of quality adjustment for health services in CPIs and PPIs shows that the basic problems with price indices are the same as with volume indices in other countries.

1.47 In conclusion, the basic measurement methods used for the market and for the non-market sector should be consistent. Consistency concerns in particular the aggregation method, where the same type of weights should be used to combine quantities or prices of services into volume or price indices. The main issue is to ensure that the quantities, prices, or unit costs refer to a unit of output, and not to units of inputs. Otherwise, the objective of measuring volumes of outputs cannot be achieved.

CHAPTER 2. MEASURING EDUCATION SERVICES OVER TIME

2.1 Terminology and concepts in education services

2.1.1 Target definition of education services

2.1 A useful starting point for the definition of 'education services' is the description of education services as provided by the UNESCO[11], consistent with the International Standard Classification of Education (ISCED-97), namely as "*organised and sustained communication designed to bring about learning*", where:

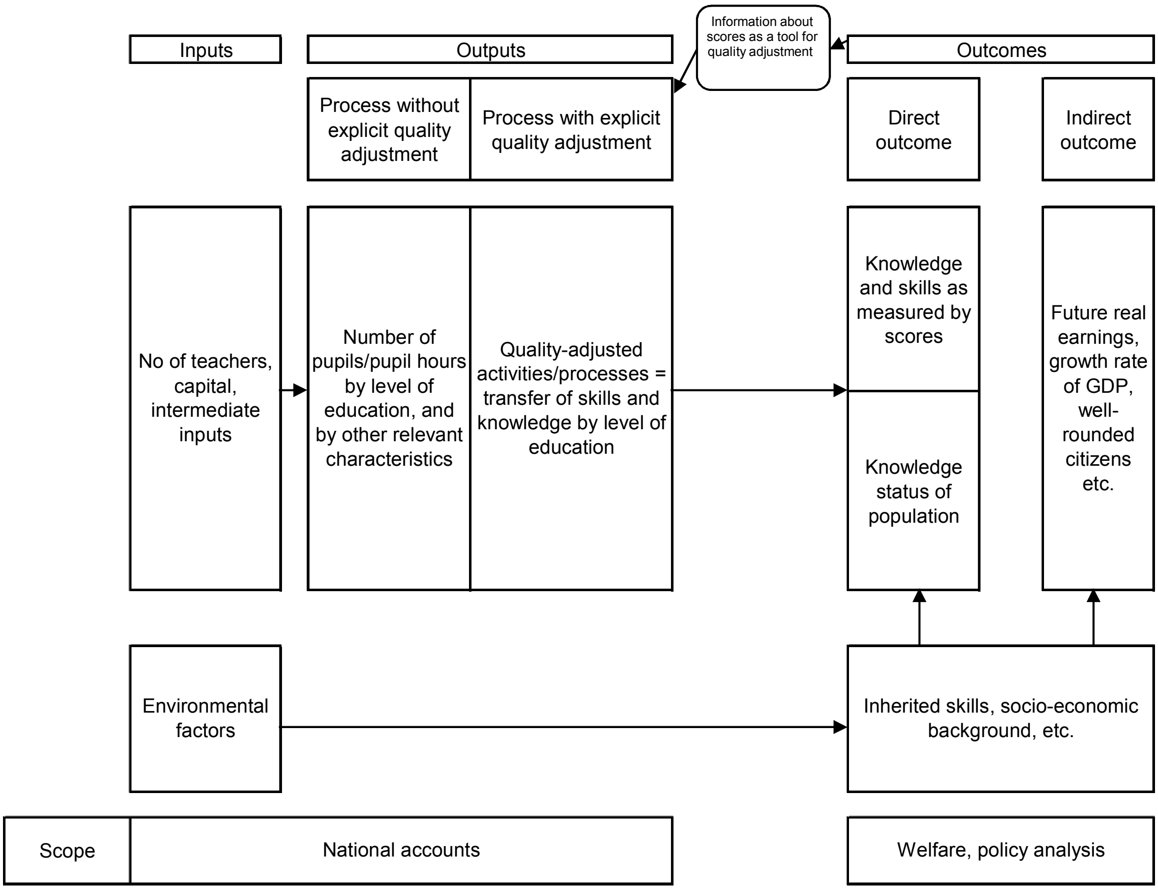
- *Communication* involves the transfer of information between two or more persons;
- *Organised* communication is that which is planned in a pattern or sequence, with established aims or curricula. It should involve an educational agency that organises the learning situation and/or teachers who are employed to consciously organise the communication;
- *Sustained* communication is that which has the elements of duration and continuity as part of the learning experience;
- *Learning* is taken as any change in behaviour, information, knowledge, understanding, attitudes, skills, or capabilities which can be retained and cannot be ascribed to physical growth or to the development of inherited behaviour patterns.

2.2 This description of education makes no direct reference to inputs or to production to define educational services; recall the differentiation between inputs, processes, output, direct and indirect outcomes. The distinction is shown in Figure 2.1 below. In the past, national accounts relied on deflated inputs: deflated wages of teachers plus deflated intermediate consumption plus deflated fixed capital consumption (depreciation of buildings for schools and universities). Because of the importance of wages in this sum, and because deflated wages are directly correlated with the number of teachers, this corresponds broadly to a method whereby the volume output is approximated by an input, the number of teachers.

2.3 For the purpose at hand, we start with the definition of learning by UNESCO and consider learning as an outcome (the change in behaviour, information etc.), and the ***organised communication of knowledge*** as the relevant output measure. The organised communication of knowledge constitutes the service provided. The individual actions which are associated with this service include activities such as preparing lessons, classroom teaching, setting examinations, marking students' work, general supervision and counselling. The organised communication is the object of transaction, explicitly so in a market context and implicitly so in a non-market context. In our definition of the output of educational services we have thus to look for variables that track the organised communication of information as defined above.

[11] UNESCO General Conference, 29th session, November 1997.

Figure 2.1. Inputs, outputs and outcomes in the provision of education services



2.4 Educational services are not homogenous because the organised communication of knowledge will vary with the level and type of education. A natural first differentiation is therefore by levels of education and these are captured by the International Standard Classification of Education (Table 2.1). Each category is defined in more detail in the source document. For example, Level 1 consists of programmes that aim to give students a sound basic education in reading, writing and mathematics along with an elementary understanding of other subjects such as history, geography, natural science, social science, art and music. It covers, in principle, six years of full-time schooling starting between the ages of 5 and 7. The rest of this chapter distinguishes between output measures for primary and secondary education and for tertiary education.

2.1.2 Activities and products

2.5 The above definition of education services is essentially a product definition. For output measurement in the national accounts, interest focuses on the economic units whose primary activity is the provision of education services. The relevant classification in the national accounts is ISIC, a classification of activities and CPC[12], a classification of products. CPC and ISCED are consistent in their definition of education services.

12 Central Product Classification, see more information on http://unstats.un.org/unsd/cr/family2.asp?Cl=16

2.6 Table 2.1 shows how the classes of activities as defined in ISIC match up with the levels of education as defined in ISCED. Several points are worth noting:

- The three-way breakdown of educational activities in (1) formal education, (2) informal education (3) support activities, characterises the main structure of any comprehensive approach of education. Formal education is the focus of the present handbook, and this is consistent with the "education data" structured and collected internationally;
- Informal education, i.e., sport or recreation education, cultural education and other education are in principal part of education activity. However, they are not specifically treated in this Handbook, although the basic principles for measuring formal education should apply;
- Educational support activities were not part of education in earlier activity classifications. In ISIC Rev. 4 this class includes: provision of non-instructional services that support educational processes or systems: · educational consulting· educational guidance counseling services;· educational testing evaluation services; educational testing services; organization of student exchange programs Overall, the size of these activities is likely to be small in comparison and no specific guidance on their price-volume split is provided here;
- The comparison between ISIC and ISCED should not stop here. A major distinction is that in an activity classification, a unit is allocated to the ISIC class that characterises its primary activity. It is frequent that units have secondary activities as well. An important secondary activity for education is, for example, the provision of research services. Output measures of the education sector as an activity should reflect the primary as well as secondary and tertiary activities, i.e., the whole set of products that is produced in the industry;
- For completeness, other relevant classifications should be mentioned. In the Classification of the Functions of Government[13] (COFOG), and in the Classification of the Purposes of Non-Profit Institutions serving Households (COPNI), the concept of education is broader than education activities and products in national accounts with research and administration regulating the activities of education included in 'education' in COFOG and COPNI.

Table 2.1. Scope and classification of education according to ISIC and ISCED-97

	ISIC rev 4 classes		ISCED-97 levels of education
Formal	8510	Pre-primary and primary education	Levels 0 and 1
	8521	General secondary education	Levels 2 and 3 oriented general
	8522	Technical and vocational secondary education	Levels 2 and 3 oriented vocational and technical
	8530	Higher education	Levels 4, 5 and 6
Informal	8541	Sports and recreation education	Not classified in ISCED-97 levels of education
	8542	Cultural education	
	8549	Other education n.e.c.	
Support	8550	Educational support activities	Not explicitly mentioned in ISCED-97 levels of education

2.7 This handbook focuses on formal education corresponding to the ISIC activities 8510 to 8530, which are also the levels of education shown in ISCED-97.

13 See http://unstats.un.org/unsd/cr/family2.asp?Cl=4.

2.1.3 A more detailed list of education services

2.8 In addition to the broad levels of education as spelled out in ISCED, it may be necessary to distinguish distinctly different education services, for example for education services for disabled students. The stratification used should be tailored to reflect, as much as possible, product differences that can subsequently be associated with their prices (in the case of market producers) or with their unit costs (in the case of non-market producers). Those education services that can be reasonably considered substitutes from a consumer perspective should be grouped together. Note that 'product difference' relates to different education services from a consumer perspective, and not to different types of inputs to produce them. Prices and unit costs will be used as weights when aggregating quantities of education services. This means that a student-hour of upper secondary education will receive a bigger weight than a pupil-hour of primary education, if the former costs more than the latter.

2.9 Table 2.2 indicates a minimal stratification, reflecting the three broad areas of education services in line with the ISCED-97 levels of education. The table also indicates a more detailed stratification that is better suited to deal with the heterogeneity of education services within each level of education. Of course, the more detailed grouping is also more data demanding.

2.10 The additional sub-stratification for normal and "special" classes "for physically or mentally handicapped pupils", in primary and lower secondary education is justified by the importance of a compulsory ancillary activity in the second case (social services), involving more costs (non educational staff, special equipments). If disabled pupils are taught in normal classes, the same idea of extra social services provided to them can be reflected by weighting the number of disabled pupils by an estimate of their relative unit costs.

2.11 The additional sub-stratification between general and pre-vocational or pre-technical education on the one hand, and vocational or technical programmes on the other hand, within the upper secondary education and the post-secondary non tertiary level of education is justified by their different purposes: general knowledge and skills in the first case, specific knowledge and skills oriented towards a specific employment in the third case.

2.12 Tertiary education has to be broken down into several sub-strata. The first consideration with regards to sub-stratification is the heterogeneity of institutions and curricula in their provision of education services. The second consideration is that there is an important secondary activity in most tertiary education units, namely the production of research, which deserves treatment as a separate product.

2.13 With regard to the first consideration, it is common practice in international work to group together "first stage: more theoretically based programmes (5A)" and "second stage: advanced research qualification (6)" and to treat this as a product distinct from "first stage: more practical and occupationally specific programmes (5B)". Inside the first category, the Bologna process (1999) has defined for European countries an equivalence of degrees which distinguishes a first cycle (ending with a bachelor's degree), a second cycle (ending with a master's degree or a doctorate). This is slightly different from the distinction between first stage/second stage in ISCED-97. The Eurostat (2001) handbook recommends a stratification that breaks university education services down by fields of education.

2.14 There are thus several possibilities for stratifying tertiary education services. The proposal here is to keep a distinction between broad fields of study such as arts, engineering, medicine etc. on the basis that these constitute distinct products. A cross-classification by level of degree may be desirable.

2.15 Concerning the second consideration, we note that research is an important secondary activity in higher education. Quantifying and quality-adjusting research output is a difficult matter that raises some of the general problems associated with the measurement of R&D activities such as whether unsuccessful research constitutes output or not. The 2008 System of National Accounts foresees, at least over a medium-term horizon, the treatment of R&D activity as the production of an investment good. Work is underway at the OECD and at Eurostat to develop practical guidance as to how to measure flows and stocks of knowledge, and this includes the issue of how to calculate volume measures of the investment good 'research'[14]. Consequently, the present Handbook does not deal with this issue.

Table 2.2. Minimum and preferred stratifications

Formal education

	Minimum stratification	Preferred stratification
Level 0	Pre-primary education	All classes
Level 1	Primary education or first stage of basic education	General classes
		Special classes, e.g. for disabled pupils
Level 2	Lower secondary or second stage of basic education	General classes
		Special classes, e.g., for disabled pupils
Level 3	Upper secondary education	General + pre-vocational
		Vocational
Level 4	Post-secondary non-tertiary education	General / vocational if available
Level 5B	More practical and occupation-specific programmes tertiary education	All classes or by professional purpose
Level 5A + 6	More theoretically-based programmes tertiary education	By fields of education and/or type of education unit, or by equivalences of degrees
Adult and other informal education	Adult and other education, anticipating extension of "education" content in ISIC rev 4, class 8540.	Adult general education
		Adult vocational education
		Computer training
		Driving lessons
		Music lessons
		Other cultural and artistic lessons
		Sport lessons
		Recreational lessons
Education support activities	According to what will be retained in class 8550 of ISIC rev.4	Other education activities

2.2 Measuring the volume of education services

2.2.1 Primary and secondary education

2.16 The system of national accounts considers primary and secondary education services as products that are consumed by households. Treating teaching services as a product consumed by households means that these services are consumed at the same time as they are produced. One implication is that grades,

14 See OECD 2009: Handbook on Deriving Capital Measures of Intellectual Property Products

credits or exam scores attained by students are not *as such* educational services. Moreover, they are largely based on learning efforts made by students and, in the SNA learning is outside the production boundary. That said, if the assumption is made that the input of student efforts is in constant proportion to teaching services provided by the educational unit, and if teaching quality is constant, then the number of degrees obtained can be taken as a proxy for the volume of teaching services. Further, and despite the fact that student attainments are not outputs as such, they reflect, at least in part, the quality of teaching. Thus, one way of gauging changes in the quality of teaching services is by using information on student attainments. This aspect will be further discussed below.

2.17 The Eurostat (2001) Handbook on Price and Volume Measures in National Accounts makes detailed recommendations on how to measure output of education services. It cites two features:

- Education is an 'individual' service, delivered to pupils by educational establishments using the types of inputs described above;

- Teaching is generally undertaken in groups of different sizes.

2.18 This begs the question whether a measure of teaching services should reflect the volume of teaching which takes place, in which case giving a lesson provides the unit of output, or whether the measure should reflect the number of pupils who are taught, in which case teaching a pupil is the unit of output? Results can be different when the size of the audience taught changes.

2.19 The Eurostat handbook concludes that the output measure should aim to reflect the sum of the individual benefit provided to each pupil (paragraph 4.12). This is consistent with the idea that the provision of services should be counted with reference to a transaction between the pupil and the service provider. In the absence of pupils, and therefore in the absence of transactions, there is no output – teaching in front of an empty classroom confers no benefits. The recommendation is therefore that a single unit of education services should be expressed as an hour of teaching received by a student at a particular level and programme orientation ('pupil hours'). Aggregate education output in a country during a period of time can be expressed in terms of the number of pupil hours at each level. But it needs to be kept in mind that this is a proxy measure for a broader notion of teaching services received and the approximation will only be valid if pupil hours and teaching services more broadly defined move in parallel and if the quality of teaching remains unchanged.

2.20 National accounts treat the output of education units as a service consumed by pupils. This is not the only way of thinking about these services. In particular, the output of the education sector can also be considered an investment good. Under such an approach (Fraumeni and Jorgenson 1992), the educational institutions, jointly with pupils, produce additions to the human capital of these pupils. Note that these investment goods are the *joint* output of schools or universities and pupils. Thus, the output measures proposed refer to output of the education *sector* at large, and not only to the output of education units. Treating students' time as input into the production process is at variance with present conventions in the national accounts where such own-account services by households are ignored. A full implementation of the human capital approach would thus entail more fundamental changes to the accounts than just a different volume measure of output unless there is a possibility to identify the marginal contribution of education services to the change in human capital. This takes nothing away from the usefulness of the human capital approach for analytical purposes or as an approach to be followed in the context of satellite accounts. The human capital approach also opens one avenue towards capturing aspects of quality change of education services, to be discussed later on.

2.2.2 Tertiary education

2.21 While the same concepts apply, tertiary education is organised differently from primary and secondary education. First, the number of lessons provided to students is more limited. This makes pupil hours not a useful measure of outputs. Second, compared to lower level education, attainments in tertiary education depend more on a student's own efforts. The higher the degree, the more characteristic this feature tends to be. Therefore, it is very difficult to clearly separate services provided by the education establishment and their quality from the input provided by the student him or herself.

2.22 A widely-used variable to measure output is the number of students. A minimum requirement is that such a measure should be based on full-time-equivalent students. Many problems remain, however. Changes in education quality cannot be captured and student numbers are also an inaccurate proxy for the quantity of services received. Participation in studies varies significantly reflecting in the length of studies. Sometimes students may have even finished their studies but prefer to stay on the register of university due to tough labour markets or merely to keep their student status.

2.23 Another approach uses the number of degrees obtained. Problems associated with this measure include ensuring the comparability of degrees over time, obtaining estimates of the relative values of various degrees, dealing with double diplomas from a single curriculum (especially with international curricula), and dealing with the fact that the contribution of the university is not separated from the education status of students that enter university and from the efforts made by the students to obtain the degree.

2.24 There is also the timing of production. It normally takes several years to graduate and in principle, production should be allocated to all years of studies rather than to the year of graduation. In this way output would also match with production costs. The simple use of changes in the number of degrees awarded as a measure of volume output can result in a bias when the number of students entering a faculty changes due to education policy or for other reasons.

2.25 Another possibility is to use data on student credits, where students' attainments during a year are measured as credits obtained from mastering various programmes during a year. Many European countries apply the European Credit Transfer and Accumulation System (ECTS) that is a systematic way of describing an educational programme by attaching credits to its components. The definition of credits may be based on different parameters, such as student workload, learning outcomes and contact hours. Credits are allocated to all educational components of a study programme (such as modules, courses, placements, dissertation work, etc.) and reflect the quantity of work each component requires to achieve its specific objectives or learning outcomes in relation to the total quantity of work necessary to complete a full year of study successfully.

2.26 Lastly, where market and non-market educational services are provided, it is possible to use the price index for market tertiary education and apply it to the value of output (i.e., to total costs) of non-market education units. To be valid, the assumption has to be made that the educational service and its change over time in market and in non-market units is at least roughly comparable.

Box 4. Is a volume measure based on processes better than a volume measure based on inputs?

Few people would disagree that a well-developed, quality-adjusted output measure of education is preferable over a volume measure of education services that relies on deflated costs or on the volume of inputs. An important question is, however, whether a volume measure of output without explicit quality adjustment, i.e., one that relies on processes, is necessarily an improvement over an input-based measure. Education provides some interesting examples here, although a similar reasoning could be applied to the measurement of health services. The following examples have been drawn from a document by Statistics Denmark.

Let us assume that there is a change in the composition of output between the different strata of education services where the latter have been defined to reflect the levels of education. If there is a change in the composition of students within a particular stratum, for example, there is a larger share of university students in the field of arts and a smaller share in the field of science, a simple volume measure of the number of university students will not change whereas a volume measure based on a detailed input method will reflect this change if costs differ.

Another example concerns changes in the input composition, for example an increase in intermediate consumption combined with an unchanged number and distribution of students. This would be the case if more and better educational materials were being used. No increase in output would be reflected in the process-based volume indicator but an input method would record an increase. Statistics Denmark concludes that this constitutes another example for a case where an input method seems to be more appropriate than the volume indicator method.

There is no final answer to this issue but the above examples highlight the importance of incorporating quality adjustments as quickly as possible, lest there be a biased measure of output. However, they also demonstrate the importance of the right level of stratification. For example, if tertiary education is further broken down by fields of study rather than lumping students of all fields into one stratum, an important step towards quality adjustment is already undertaken and the above criticism of the output measure would no longer apply.

2.27 In summary, as a first step towards measuring education output (without explicit quality adjustment), the present Handbook recommends for primary and secondary education services the number of pupil-hours, differentiated by the level of education and possibly other characteristics. We note again that services received by pupils consist of wider set of services as personal instructions etc. but nevertheless pupil-hours appears to be a satisfactory proxy for output in situations where teaching in classrooms is a dominating channel in transferring knowledge. No clear recommendation emerges for tertiary education. Among the measures considered, the application of a price index for market-provided tertiary education services to non-market services constitutes a reasonable approach provided such an index is available.

2.2.3 Quality adjustment – general considerations

2.28 Measuring outputs along a timeline poses challenges. It is relatively easy to measure when more of the same type of output is produced between two periods: we could be comparing 200 hours of teaching at a particular level received by students this year with 100 hours the previous year, and hence note a doubling of output. However, the nature of the education service can change over time. That makes comparisons – and estimates of growth – more challenging as like is no longer compared with like.

2.29 It seems fair to say that the most important quality characteristic of teaching services is how they succeed in advancing knowledge and skills of students. But other characteristics of services valued by consumers can also shape quality, including those that are not directly related to the transfer of knowledge and skills such as working atmosphere, safety etc.

2.30 In addition to "core educational services" there are other, non-educational goods and services produced by education units, for example catering, boarding services, and transport services. To ensure comparability of education services over time, these services should be deflated or directly estimated in volume separate from education services and consistent with the kind of product provided (see also Eurostat 2001).

2.31 Broadly, there are three ways of accounting for changing quality in education services:

- Stratification: different types of education require typically different types of teaching services. When the number of pupil hours are compared at a sufficiently detailed level to identify different educational products, this stratification constitutes a form of implicit quality adjustment – like is compared with like over time and in space;

- Explicit quality adjustment: when the implicit control for quality change via stratification is insufficient, the final output measure may require an *explicit* quality adjustment. Explicit adjustments require identification of quality characteristics that are not present in the stratification process and the calculation of an adjustment factor that is applied to unadjusted measures of teaching services. Such adjustments could for example reflect exam scores (O'Mahony and Stevens 2003). How this can be done is described in greater detail below. Suffice it to say here that one major challenge lies in separating those effects that are attributable to the education system from those quality changes that are due to other factors, unrelated to the provision of education services. Another difficulty lies in translating units of scores into units of output. A variant of the explicit quality adjustment is to use outcome measures directly as measures of output. For instance, the number of exam scores could be taken as the output of the education sector, provided that it is possible to capture only those movements in exam scores that are due to teaching services;

- Use of indirect outcome measures: the human capital approach towards measuring education services relies on information from *indirect* outcomes (future earnings due to education) to capture an element of quality change. Another approach towards quantifying quality differences in educational services is the housing value approach that relies on the information about house price differences for houses near borders of school districts with differential performance ratings (Black 1998).

2.2.4 Capturing quality change in secondary education

2.32 There are several approaches towards quality-adjusting measures of educational services. Quality of teaching is typically multi-dimensional and consists of a number of factors valued by consumers. For primary and secondary levels, the most common adjustments are based on school inspection results, exam scores, and class-size, all taken to reflect the quality of teaching services in one way or another. Consider them in turn.

School inspection

2.33 School inspection results provide useful information by identifying characteristics of education services that are valued by users and are thus valid quality indicators. These characteristics could provide guidance concerning the stratification of education services, but their multi-dimensional and often qualitative nature makes it difficult to use them for explicit quality adjustment. The indicators do not normally follow a metric scale as required for the estimation of volumes and even if they did it is not clear what kind of weighting should be applied in the aggregation of various indicators. Also, the indicators are typically not mutually independent. For these reasons, subjective elements remain with adjustments based on school inspection results.

Class size

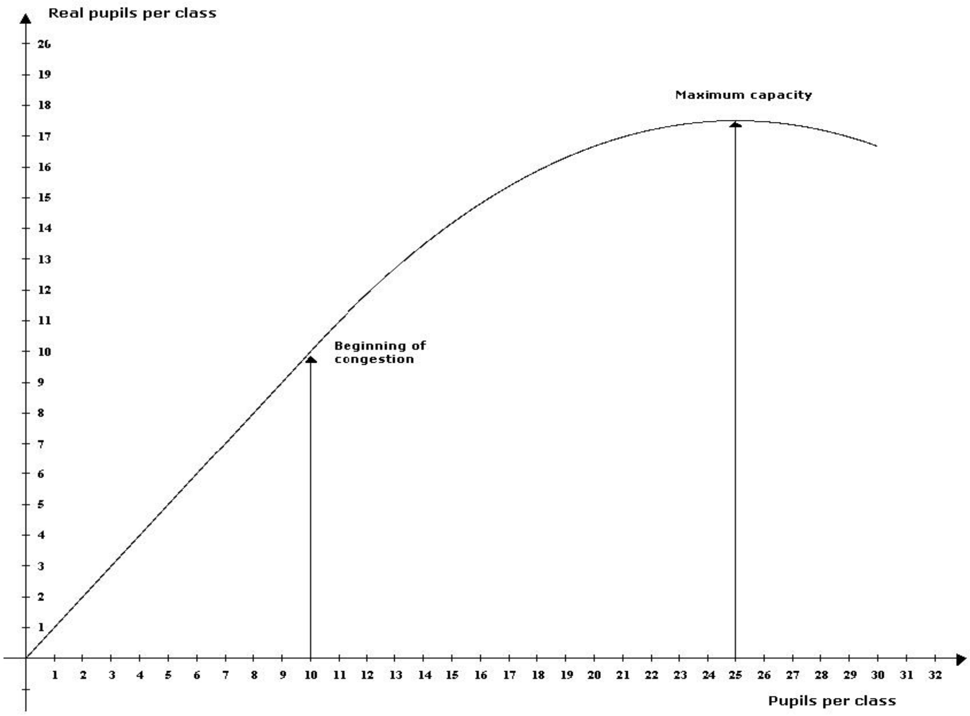
2.34 Another, frequently-quoted characteristic of teaching services is class size. It has been argued that the quality of teaching improves with a decreasing class size because the teacher has more time for each pupil. There is a large and varied literature on the topic and the references below are necessarily incomplete. Also, it is not always clear to what extent studies control for effects other than class size. Thus, conclusions have to be taken with the necessary caution.

2.35 Christian (2006) provides an example from the United States where the pupil-teacher ratio in public elementary and secondary schools declined from 18.7 to 15.9 between 1980 and 2001 suggesting that the quality of the education provided improved over this period. To translate changes in class size into coefficients for quality adjustment, Christian used two different research results. The study by Rivkin et al. (2005) concluded that class size was negatively correlated with the educational achievement of pupils: classes with fewer pupils show higher total achievement. More specifically, a reduction of the number of pupils by 1% resulted in an improvement of 0.001 standard deviations in mathematics test scores per year. Christian's second reference was the study by Bowles, Gintis and Osborne (2001), in which the standard deviation of test scores is equivalent to one year of education. Christian concluded that a reduction in the number of pupils results in 0.1% quality improvement of teaching. At the same time, an analysis of the National Assessment of Educational Progress (NAEP) used mathematics test scores and showed that one standard deviation is the equivalent of 3.3 years of schooling and, thus, a reduction of class size by 1% results in a much larger quality improvement of 0.33% (Christian, 2006).

2.36 The above adjustment enters in a linear fashion. It can be argued, however, that class size does not influence quality significantly as long as classes are small but deteriorates if classes become congested. Collessi et al. (2007) propose the following model, which assumes that congestion in a class begins with 10 pupils. Complete congestion is reached at class size 25: if 25 or more pupils are taught by the same teacher, this is equivalent to only 18 'real' or 'fully-taught' pupils, i.e., children taught at normal conditions.

2.37 With this non-linear relationship between class size and 'fully-taught' equivalent students, the distribution of class sizes in the country enters the picture: if a national average of 20 pupils/teachers reflects many classes of 10 pupils and many classes of 30, this yields a different 'fully-taught' equivalent number than a situation where the national average reflects a majority of classes of 20 pupils. The right use of this quality indicator is to convert, for each class, the "number of enrolled pupils" into a 'fully-taught'-equivalent number, which constitutes the quality-adjusted measure of output.

Figure 2.2. The 'outcome-oriented' Italian model for class size



2.38 A way to be considered is also to estimate the influence of class size indirectly by applying a detailed stratification by class size and then measuring pupil-hours for each stratum. In this case, for every class-size category the change in the number of pupils would be driven by the change in the number of classes and, provided that required teaching hours per class remain unchanged, the change equals also the number of teachers. Thus, in this case results become very similar to results derived by using the traditional input approach (Box 5).

2.39 Overall, while there is significant literature on the relation between class size and the quality of teaching, empirical results do not provide a uniform answer. A reference study would thus have to be chosen at the national level. Also, periodic updates of such a study would be necessary to ensure that results remain relevant. But experience in some countries for instance Italy shows that quality adjustment on the basis of class-size effects is a feasible approach.

Box 5. A finer stratification may give similar results to an input method

While detailed stratification is useful, it has to be applied with care to avoid counter-intuitive results. Box 1 already presented an example where detailed stratification can be problematic if different medical treatments with the same result for consumers are treated as separate products. The same idea applies to education. A conceptually correct output approach would reason in terms of different products supplied to consumers, not different processes employed in producing them.

An interesting example relates to stratification by class size, the idea being that each class size defines a different product. For simplicity say that the costs are wages for teachers, and that a single wage rate applies to all teachers. For each stratum then, basic quantity change would be measured as the change in the number of pupils attending a class of size *i*, $Pupils_i^t/Pupils_i^{t-1}$. The volume index is the weighted average of these quantity indices, the weights being the cost shares of each type of class size. With the simplifying assumption that teachers all earn the same wage, the cost share equals the share of teachers:

$$I = \sum_{i=1}^{N} \frac{L_i^{t-1}}{\sum_{i=1}^{N} L_i^{t-1}} \left(\frac{Pupils_i^t}{Pupils_i^{t-1}} \right)$$

As the strata were defined by class size, the number of teachers and the number of pupils are proportional, that is $L_i^t \cdot i = Pupils_i^t$ Inserting this relationship into the definition of the volume index above yields

$$I = \sum_{i=1}^{N} \frac{L_i^{t-1}}{\sum_{i=1}^{N} L_i^{t-1}} \left(\frac{L_i^t}{L_i^{t-1}} \right) = \frac{\sum_{i=1}^{N} L_i^t}{\sum_{i=1}^{N} L_i^{t-1}}$$

Thus the volume index obtained by this "process-based stratification" is the change in the number of teachers, i.e. the same volume index as in the traditional input method. Note, however, that the example is based on simplified assumptions. Firstly, it is implicitly assumed that teaching hours per teacher are constant over time, which is not necessarily true. Secondly, costs, including teacher salaries, are not necessarily independent of class size. But, third, *if* it is actually true that class size is the most important quality characteristic, output *should* vary with the number of teachers.

Exam scores

2.40 The most frequently invoked explicit quality adjustment to pupil hours or to the number of pupils taught is on the basis of exam scores. The applicability of this method of quality adjustment will depend on the level of education. For example, there is little point in recommending the use of exam scores for the quality adjustment of pre-primary education. For tertiary education, several possibilities exist, including the use of scores and the use of future real earnings. For secondary education, an important quality component of the output is how much a school can be expected to contribute in attaining knowledge and skills, and data on exam scores may provide a conceptually correct and empirically feasible option for explicit quality adjustment. This aspect will be explored more closely here.

2.41 If one takes as a starting point that changes in pupil attainments are the primary indicator of the quality of educational services, one way to capture quality is to assess how they, on average, develop over time. Other quality components exist, but to keep things tractable, an assumption has to be made that for a given level of education, the single most important quality criterion is the transfer of knowledge because this permits expressing educational output in a single, measurable dimension. Also, education policy targets such as even distribution of skills are not considered when focussing on average scores.

2.42 In the case where educational outcome can be assessed by a score, the quality adjustment of the output measure should be proportional to the change in the score brought about by teaching. There are two issues here: how to establish a factor of proportionality between the change in scores and the change in output – a 10% rise in scores does not automatically translate into a 10% rise in the volume of output. The second issue relates to the fact that changes in scores have to be limited to the marginal contribution of the education service.

2.43 *Contribution* of the education unit means that only that part that is attributable to the education unit in the average change in the education outcome of a pupil should be measured for quality adjustment. Other contributions should be excluded, in particular contributions by:

- The pupil him/herself (inherited skills or special efforts, motivation);
- The family (socio-economic background, intellectual stimulation by parents etc.);
- Another education unit (extra school remedial courses, private tutoring if the focus is on a unit and not on the whole education system, etc.).

2.44 These other contributions to the pupils' level of knowledge and skills explain mainly why different pupils obtain different outcomes while receiving the same teaching (hence, the same output). The significance of the non-teaching contribution to attainment of results depends on whether the same pupils are tested over time or whether the test is made only for pupils at a certain age or stage of education (e.g. PISA). If tests are made at the beginning and end of the period, they are not necessarily comparable between different ages. Nevertheless, some studies exist where results from exams taken with pupils of different age are used to establish a quantitative link between scores and years of education (see Box).

2.45 The target measure for the quality-adjusted volume change of education services is the change in the number of pupil hours (H) multiplied by the quality of teaching. The indicator for the quality of teaching is average scores (S) divided by the change of pupil hours per pupil (H/N). Division by H/N is necessary because pupil attainments are influenced by possible changes in the number of lessons and this influence should be eliminated to arrive at quality of one pupil hour. This leads to the following formula:

$$\text{Change in the volume of education services} = \frac{H^t}{H^{t-1}} \times \frac{\frac{S^t}{S^{t-1}}}{\frac{H^t/N^t}{H^{t-1}/N^{t-1}}} = \frac{N^t}{N^{t-1}} \times \frac{S^t}{S^{t-1}}$$

2.46 Thus the formula reduces to a form where education output is simply the number of pupils multiplied by average scores. Thus, the volume equals the successful transfer of knowledge and skills to pupils - provided that the test has been well-designed.

2.47 The formula is also consistent with Eurostat (2001). If indicators of pupil attainments are not available and no quality adjustment can be undertaken, the formula provides the number of pupil-hours, which is the first best quantity indicator recommended by the European handbook. If neither indicators of pupil attainments are available nor data on pupil hours, the formula reduces to the number of pupils, which is the second best quantity indicator recommended by the European handbook.

Box 6. Linking exam results and years of education

Christian (2006) reports that the "Analysis of National Assessment of Educational Progress (NAEP) math test scores suggests that a standard deviation of test scores is the equivalent of 3.3 years of schooling (the standard deviation of math NAEP scores for 17-year-olds is about 31 points; this was approximated by observing the percentile distribution of scores in 1996 and assuming a normal distribution. The average math NAEP score improved from 231 at age 9 to 307 at age 17. Dividing the difference between these two scores by 8 yields an annual NAEP gain of 9.5 points, which is about 1/3.3 the cross-sectional standard deviation of 31)."

To illustrate this example, suppose there are two schools, A and B, with 100 pupils per grade between the age of 9 and 17. In school A, students achieve average scores of 231 for 9-year old pupils – the entry – and an average score of 307 for 17 year olds at the end. In school B, 9-year old students score the same average as in school A, i.e., 231 points. Thus, at school entry, both student populations show the same characteristics. However, in the final scores, the 17-year old students in school B obtain an average result of 338 points, which corresponds to 307 in school A plus one standard deviation, i.e., 307+31=338 points. The annual average change of scores (required adjustment rate) is 3.6% for A and 4.9% for B.

2.48 A more typical situation is that pupils are tested at certain age or stage of education. Results can then be compared with results for pupils that participated in the test in the previous period or periods. An example is the OECD PISA assessment[15] that is run every three years. In this case comparable tests are easier to design but, because the target group is always different, more attention should be devoted to the elimination of effects that are not attributable to teaching.

2.49 The environmental factors of consumers, also called non-discretionary inputs, should thus be neutralized in an estimate of output. In education, it means that the family background and even the inherited skills have to be subtracted from the education outcome to provide an estimate of the output. For example, PISA undertakes an "Economic, Social, and Cultural Status (ESCS)" correction where results from every country are modified to show those values that would have been observed had every country had the same ESCS structure as the OECD on average. Other examples how environmental variables can be controlled for can be found in national comparisons of education units: comparisons are made with outcomes that account for the economic and social characteristics of the pupils, but the correction for environmental factors is often forgotten when time-series of exam scores are used for quality adjustment. The economic, social and cultural structure moves slowly, too much to be observed with reliability each year, but this change has a strong influence over 10 or 20 years.

2.50 A problem in this kind of test is that results are not only influenced by recent teaching but all teaching provided to the pupils since their entering the school up to date when the test is made. A solution suggested by Christian (2006) is that the difference of scores is simply allocated linearly over the earlier school years.

Translating differences in scores into quality adjustments

2.51 In formal education, academic scores can usually be defined and harmonized at a fine level of education fields and grades. It is also likely that score data do not exist for all education fields and grades but score data are available for only a part of curriculum at best. For fields that are not covered by tests, use of pupil hours is possibly the only alternative.

15 For more information about PISA, see www.pisa.oecd.org.

2.52 Also, academic scores have an element of subjectivity as every teacher has a different subjective method of scoring. Ordinal rankings of pupils or the use of minimal thresholds can help making scores more "objective", but makes it difficult to quantify differences between pupils.

2.53 The first step usually consists in normalising the Standard Deviation of scores across countries or over time. It seems that many national and international assessments are standardised with two alternative conventions: either 1 Standard Deviation (SD) = 10% of the average (US NAEP test in mathematics), or 1 SD = 20% of the average (PISA, PIRLS and TIMSS, all international tests). If we suppose that the distribution of scores respects a "normal" frame, the 16th centile of pupils performs the average – 1 SD, while the 84th centile of pupils scores the average + 1 SD. This means that in the case of 1 SD = 20%, the second category is supposed to know (1 + 0.2) / (1 - 0.2) = 1.5 times more than the first category, or in the case of 1 SD = 10%, the second category is (1+0.1) / (1–0.1) = 1.22 times more skilled than the first one. International or temporal comparisons will therefore require using the same norm. Scores realised under a particular SD can be approximated in another SD by simple mathematical transformation.

2.54 There is some evidence that the normalisation of the observed distributions to a common mean and to a SD of 10% or 20% of the mean may be a reasonable choice. In the economic literature, returns to education have been used to link exam scores and future real earnings. Krueger (2003) reviewed three studies:

> "Muriane, Wilet and Levy (1995) estimate that male high school seniors who scored one standard deviation (SD) higher on the basic math achievement test in 1980 earned 7.7 percent higher earnings six years later, based on data from the High School and Beyond survey. The comparable figure for females was 10.9 percent.
>
> Currie and Thomas (1999) use the British National Child Development Study to examine the relationship between math and reading test scores at age 7 and earnings at age 33. They find that students who score in the upper quartile of the reading exam earn 20 percent more than students who score in the lower quartile of the exam, while students in the top quartile of the math exam earn another 19 percent. Assuming normality, the average student in the top quartile scores about 2,5 standard deviations higher than the average student in the bottom quartile, so their results imply than one SD increase in reading test performance is associated with 8,0 percent higher earnings, while a one standard deviation increase in the math test is associated with 7,6 percent higher earnings.
>
> Neal and Johnson (1996) use the National Longitudinal Survey of Youth to estimate the effect of students' scores on the Armed Forces Qualification Test (AFQT) taken at ages 15-18 (adjusted for age when the test was taken) on their earnings at age 26-29. They find that a one SD increase in scores is associated with about 20 percent higher earnings for both men and women".

2.55 The above studies would suggest that one SD in scores corresponds to between 8% and 20% in earning differences, depending on the study. If relative earnings reflect relative skills and knowledge, this provides some support for the normalisation procedure of the distribution of exam scores mentioned above which sets the standard deviation to 10% or 20 % of the mean, depending on whether one looks at NAEP or at one of the international student survey results. There is a large margin between 10 and 20 % and no objective way of choosing one over the other, but these two measures provide a benchmark. We conclude provisionally that academic scales for direct outcomes are necessarily subjective, but there is some foundation for standardising to an SD that is 10-20% of the average score.

2.3 Quality adjustment with real future earnings

2.56 An alternative approach towards capturing quality change in the provision of education services has been pioneered by Jorgenson and Fraumeni (1992). The approach which can be applied to a broad range of educational levels draws on the literature on human capital. Jorgenson and Fraumeni (1992) describe their approach as follows:

> "…economists have found it useful to characterize the benefits of education by means of the notion of investment in human capital. This idea captures the fact that investment in human beings, like investment in tangible forms of capital such as buildings and industrial equipment, generates a stream of future benefits. Education is regarded as an investment in human capital, because benefits accrue to an educated individual over a lifetime of activities."

2.57 When education is treated as an act of investment in human capital, traditional capital theory can be applied which suggests that the value of a capital good equals the discounted stream of future benefits associated with this capital good. In the case of education, the value of investment in one additional year of education is evaluated by discounting the stream of increments in lifetime income that is associated with the additional year of education. Because increments are not constant across years of schooling, this calculation has to be carried out separately for each level of education. Also, market and non-market labour income have to be considered.

2.58 Given the incremental value of an additional year of education, the total value of investment in education is measured by multiplying this incremental value by the number of pupils enrolled at a given level of education. The human capital approach then computes a volume index of educational services by weighting the rates of change of pupil numbers by the share that each group of pupils occupies in the total value of investment in education. In the simplest case of a Laspeyres volume index, one can also say that the volume of education is measured as the constant-price value of educational investment.

2.59 How does the human capital approach relate to the general framework outlined at the beginning of this chapter and more particularly, to the volume index on the basis of quality-adjustment with score results? In the human capital approach, the quantity element for a given level of education is the number of pupils. Quality of educational services is captured but only indirectly when relative 'prices' of a particular level of education change over more than one time period, thus modifying the weights in a volume index of education. In the first approach the quantity element directly includes the quality adjustment. The human capital approach implies quality adjustment that relies on the contribution of education to indirect outcomes, whereas an explicit adjustment for quality on the basis of exam scores relies on the contribution of education to direct outcomes.

2.60 There are three major advantages to the human capital approach:

- The use of monetary values does away with the problem of scaling when units of exam scores have to be translated into units of educational services;

- At least for the case of market labour income, market prices provide a relatively objective basis for the weights that enter the volume index of education;

- By looking forward and using the notion of lifetime income, a good deal of the issue associated with lags between educational services and their effects can be attenuated;

- A main attraction of the human capital approach lies in its capacity to provide *value weights* for the quantity indexes of student numbers. In principle, these value weights can be used without necessarily changing the total value of output at current prices, as determined by the sum of costs. A more fundamental consideration is to which extent value weights should feature in the national accounts in the first place, and this issue has been discussed in Section 1.2.4.

2.61 There are, however, also a number of drawbacks to the human capital approach:

- The approach followed by Jorgenson and Fraumeni (1992) considers in fact the output of the education sector at large where educational units and students *jointly* produce the investment good 'addition to human capital'. In this context, students' time spent studying is considered an input to the education sector. While this is analytically attractive, it is at odds with national accounts conventions that consider own-account production of household services to lie outside the SNA production boundary. And the measure of output is not directly usable as a measure of the output of educational institutions as understood in the national accounts;

- There may be a selection bias in the sense that individuals with above-average ability or socio-economic background are the ones who both go through additional levels of education and are also successful on the labour market. The implication being that relative earnings reflect the individuals' personal capacity and background rather than a contribution from the education system. With sufficiently detailed data, for example from censuses, it may, however, be possible to control at least for some socio-economic effects in relative earnings;

- The link between education services and future returns on education may be tenuous for other reasons as well: the structure of earnings may reflect trends in labour market that are independent of qualifications of individuals and human capital investments do not in general cease with the termination of schooling, because there is on-the-job and vocational training. The point was also clearly recognised by Jorgenson and Fraumeni (1992) who state that "… investment in education is only one of many forms of investment in human capital. Important investments are made by families in the rearing of their children and by employers and workers in on-the-job training";

- Despite the fact that relative earnings constitute an objective measure, the approach requires a number of assumptions, notably about the overall rise in expected real income, and about the rate of return for discounted flows of income. If non-market income is to be included in life-time income, this raises a series of measurement challenges;

- The human capital approach provides independent estimates of the value of education services. This poses a problem for national accounts conventions where the value non-market production is conventionally measured as the sum of costs. If the value of non-market education services is different from its costs, any treatment inside the national accounts would have to deal with the surplus or deficit item so measured. Also, the fact that the human capital approach considers education output an investment good would imply including human capital within the asset boundary of the national accounts. This would constitute an important change to present accounting conventions;

- A difficulty in the use of the human capital approach is that the relationship between an additional year of studies and change in discounted future earnings is not linear. "Normal" years of studies have much less importance for future earnings than years when degrees are achieved. To maintain, a link to actual provision of services, assumptions should be made about the share of students that will finalise their exams.

2.62 In summary, the human capital approach is a valid alternative to other measures of education output, provided there is a good empirical base for estimating current and expected returns to education. Its biggest drawback from a national accounts perspective is that in its original form it does not comply with the production boundaries set in the SNA. Rather, it captures the output of the educational sector at large, i.e., including students' efforts.

2.4 Summary of measurement proposals

2.63 For most education services up to and including secondary education, the basic indicator recommended here is the number of pupil hours, suitably stratified by level and orientation of education. Nonetheless, this measure remains a proxy to a more elaborate measure of pupil numbers adjusted for changes in educational attainment that are attributable to educational services. Quality adjustment is a central step towards the measurement of the volume of education services. The table below summarises this Handbook's proposals on output-based measures.

Table 2.3. Overview of indicators for volume output of education service providers

<table>
<tr><th></th><th>Output-based methods
Methods with an asterisk* are proxy methods</th></tr>
<tr><td></td><td></td></tr>
<tr><td>Pre-primary education</td><td>Number of pupil-hours</td></tr>
<tr><td>Primary education</td><td rowspan="10">Number of pupils, adjusted for change in pupil attainment

Number of pupil hours*

Number of pupils*</td></tr>
<tr><td>Primary education: general</td></tr>
<tr><td>Primary education: special education, e.g., for disabled pupils
Note: The sub-stratification normal / special could be replaced by coefficients reflecting the extra costs for social services provided to disabled pupils</td></tr>
<tr><td>Secondary education</td></tr>
<tr><td>Lower secondary: general</td></tr>
<tr><td>Lower secondary: special classes, e.g. for disabled pupils</td></tr>
<tr><td>Upper secondary education: general + pre-technical or pre-vocational</td></tr>
<tr><td>Upper secondary education: vocational</td></tr>
<tr><td>Post-secondary non-tertiary education</td></tr>
<tr><td>Tertiary education</td><td rowspan="3">Deflation with price index from market-based education services
Credits (ECTS)
Full-time equivalent students*
Enrolled students*</td></tr>
<tr><td>Tertiary education with practical and occupation-specific programmes</td></tr>
<tr><td>Tertiary education with more theoretically-based programmes
Note: differentiation by field of education useful</td></tr>
</table>

ANNEX 2.A: OVERVIEW OF COUNTRY PRACTICES

Results of a Eurostat/OECD survey

Introduction

2.64 In June 2006, Eurostat and the OECD asked their respective countries to provide information on the method followed for volume and price in education, including details on the stratification, the quantity and the quality indicators, if such are applied. A follow-up questionnaire with a request for updates was sent in early 2009. This section summarizes the results.

2.65 A number of OECD countries had already implemented output measures for education (Australia, Finland, France, Germany, Italy, Netherlands, New Zealand, Spain, United Kingdom) and more European countries were supposed to follow by the end of 2006. Other European countries especially new Member States, expect to apply output methods in the upcoming years (Slovenia has now implemented output-based methods), Australia and New Zealand are ahead of most other OECD countries in the implementation of output-based measures for education services. In the United States, and in Korea, these output measures are still at a research stage, without a precise timing for implementation. Canada, Japan and Switzerland have no current projects.

Stratification

2.66 The ISCED-97 levels of education were the natural reference for stratification, but some countries concerned had to aggregate some ISCED levels, either pre-primary and primary education, primary and lower secondary education, or lower secondary and upper secondary education.

2.67 Some countries use also a sub-stratification of the ISCED levels, for instance four countries distinguish the "special" classes for the purpose of disabled pupils, because of large differences in unit costs. Sometimes it turns out to be difficult to separate primary from lower secondary special education. Four countries also split the upper secondary level into general and vocational education and four countries also use a regional stratification. Two countries envisage stratification by gender, because girls and boys are reputed to have different skills in mathematics and reading.

2.68 In tertiary education, four countries use a split between university and other institutions, three countries privilege the field of education and one country (Finland) uses a combination of these two dimensions. Germany, for instance, combines a split by institutions (university and three other types of institutions) and a split by fields of education (linguistics and other cultural sciences, medicine, sports, mathematics and natural sciences).

Table 2.4. Overview of country practices in the volume measurement of education services

Country	Are output methods applied?	Main sources	Complementary stratification to ISCED 97?		Quantity indicator		Weights	Quality adjustment?
			Pre-school -> upper secondary	Tertiary	Pre-school -> upper secondary	Tertiary		
Canada	No current project							
Japan	No current project							
Korea	Research							
Luxembourg	Data since 2000	Administrative data	Ordinary / disabled + stratification between lower and upper secondary by type of school	Stratification between university and short technical superior education	Pupil hours	Number of pupils	Costs	No
Switzerland	No current project							
United States	2 research projects	Mainly surveys	Less detail than ISCED : pre-school with primary and lower secondary, perhaps upper secondary too.		Number of pupils	Number of pupils	Costs	Several possibilities : pupil-teacher ratio, high school drop-out rate, test scores,
Australia	Data since 1989	Administrative data	Special ANZICS classification ; pre-school with special, primary with secondary	Three HECS (Higher Education Contribution Scheme) bands	Number of pupils	Number of pupils (+ number of publications for research)	Costs	No
Austria	Expected by end 2009	Administrative data (public) + survey (private)	Breaking down for lower secondary between general / vocational + regional stratification	University / other + regional stratification	Number of pupils	Number of pupils	Costs	No
Belgium	Data since 1995	Administrative data	Aggregation pre-school and primary (1), lower secondary and upper secondary (2) + regional stratification	University and non university / other + regional stratification	Number pupil hours (number of pupils compiled for the sake of comparison)	Number pupil hours (number of pupils compiled for the sake of comparison	Costs. For universities research activities excluded	No
Czech Republic	Expected by end 2010	Administrative data	Ordinary / disabled ("special") in kindergarten, primary and secondary school		Number of pupils	Number of pupils	Costs	No
Denmark	To be implemented in 2012, work still going on	Administrative data	No	No	Number of pupils	Number of pupils	Costs	Rates of educated teachers, class size and scoring of the pupils

Table 2.4. Overview of country practices in the volume measurement of education services (continued)

Country	Are output methods applied?	Main sources	Complementary stratification to ISCED 97?		Quantity indicator		Weights	Quality adjustment ?
			Pre-school -> upper secondary	Tertiary	Pre-school -> upper secondary	Tertiary		
Estonia	Data since 2006	Administrative data	No	No	Number of pupils	Number of pupils	Costs	No
Finland	Data since 2000	Administrative data	Ordinary, severely disabled, other disabled (primary and lower secondary educations together) + breakdown of upper secondary between general (=<, > 18 year old), vocational (7 fields)	Polytechnic education (7) / university basic degree (20) / higher degree / research	Number of pupils + student year for voluntary further vocational training	Number of pupils, basic degrees, credit numbers	Costs	No
France	Data since 1981	Administrative data + survey	Ordinary / disabled + breakdown of upper secondary between general, vocational, apprenticeship	Short technical superior (5B) / long superior (5A + 6)	Number of pupils	Number of pupils	Costs	Pupils moving up between 1981 and 2001
Germany	Data since 1991improved from 2004	Administrative data + survey	Only special classes for disabled pupils not subdivided by ISCED. Partly regional stratification. Additional stratification from lower secondary and upper secondary by type of schools.	Stratification by institutions(universities and three other kinds) and fields (up to 9) i	Pupil hours	Number of pupils	Costs (informatio n only for public education)	No
Greece	Expected by end 2006	Survey	Aggregation lower and upper secondary		Number of pupils	Number of pupils	Costs ? (specific CPI)	No
Hungary	Data since 2001	Administrative data	Aggregation lower and upper secondary	No	Number of pupils	Number of pupils	Costs	Class size
Italy	Data since 2001	Administrative data + survey	Distinction between State and other public schools (pre-school); the upper secondary education is stratified by type (secondary school focusing on humanities, secondary school focusing on sciences, teacher-training institutes and schools, vocational institutes, art schools)	Groups of 18 homogenous faculties	Number of pupils	Number of pupils	Costs	Class size + degrees / number of students. For Universities two in-dicators: 1) ratio between number of students attending university courses within regular deadline and total students; 2) time distance from the regular length of the university courses

Table 2.4. Overview of country practices in the volume measurement of education services (continued)

Country	Are output methods applied?	Main sources	Complementary stratification to ISCED 97?		Quantity indicator		Weights	Quality adjustment?
			Pre-school -> upper secondary	Tertiary	Pre-school -> upper secondary	Tertiary		
Latvia	Under construction	Administrative data + survey	Public / private + regional stratification + age and sex ?	Public / government dependent private / independent private + regional stratification	Number of pupils	Number of pupils	Number of teachers	Class size (primary + lower secondary).
Lithuania	Expected by end 2006	Survey		Stratification by subject	Number of pupils	Number of pupils	Costs	Examination data (upper secondary and tertiary)
Malta	After 2006	Administrative data + survey	Age and sex?	Fields of study	Number of pupils	number of pupils		Pupils moving up / graduates
Netherlands	Data since 1995	Administrative data	Kind of education (early selection)	NA	Number of pupils (corrected for pupils who go to school only partially)	NA	Costs	No, except for higher education where the expected course duration in reference to Higher vocation is taken into account
New Zealand	Data since 1978	Administrative data + survey	Special ANZICS classification ; but strata less far from ISCED than Australia		Number of pupils (primary aggregated with secondary)	Number of pupils (full time equivalent)	Value added	No
Norway	Implemented	Administrative data	Owner (CGov, LGov, private) + ordinary/disabled or general (3) / vocational (13) in upper secondary	Groups of faculties having different unit costs	Pupil hours. For upper secondary, based on full time equivalent except for the private sector (number of pupils)	Number of pupils	Costs	On the agenda
Poland	Under construction	Administrative data + survey	Regional and by subject	Regional and by subject	Pupil hours	Number of pupils		Class size (primary, secondary). Use of enrolment rate In sciences and technology for tertiary education

Table 2.4. Overview of country practices in the volume measurement of education services (continued)

Country	Are output methods applied?	Main sources	Complementary stratification to ISCED 97?		Quantity indicator		Weights	Quality adjustment?
			Pre-school -> upper secondary	Tertiary	Pre-school -> upper secondary	Tertiary		
Slovakia	Expected by end 2006 (later for higher education)	Administrative data			Number of pupils and pupil hours ?		Costs	Scoring of pupils ?
Slovenia	Data since 2003	Administrative data + survey	Aggregation of primary level and lower secondary level of education + sub-stratification by educational program + classes for the purpose of disabled pupils; aggregation of upper secondary level and post-secondary, non tertiary education +sub stratification by program	By type of education program + position in the national degree/qualification structure + field of study for university.	- Pupil hours - Number of pupils	Number of pupils	Costs	Scoring of pupils and number of pupils per teacher
Spain	Data since 2000	Administrative data			Number of pupils	Number of pupils	Costs	Class size
Sweden	Data since 2002	Administrative data+survey	No	No	Pupil hours and number of pupils	Number of pupils	Costs	Use of school grades for lower secondary; number of pupils with final grade for upper secondary; for tertiary, number of students with grades
United Kingdom	Data since 1996	Administrative data	Aggregation of pre-primary (including publicly funded, private, voluntary and independent provision), primary, secondary (including city technology colleges and academies), and special schools. Stratified by component regions of the UK (England, Wales, Scotland and Northern Ireland)	Publicly funded further education for under 19s; Higher education (initial teacher training and health professional training)	Full Time Equivelent (FTE) pre-primary pupils; Attendance Adjusted FTE primary, secondary and special school pupils.	Number of students	Costs	Estimated contribution that each year of schooling makes to performance in school leaving examinations taken at age 15 -16, applied to both primary and secondary levels. Initial teacher training adjusted by proportion of final year students achieving qualified teacher status.

Quality adjustment

2.69 Of the 21 countries that use output methods, 7 make no explicit quality adjustment, and have no project for doing so. Five countries envisage to use examination data (the United States has also listed test scores among its possible quality indicators). Five countries use or have used an adjustment by the proportion of pupils moving up or being graduated, four countries consider the class size as a quality component. No country seems to have relied on school inspections.

Inclusion of non educational goods and services ("ancillary services")

2.70 The non-European OECD countries were asked about the inclusion or exclusion of the cost of all the complementary services such as canteens, cafeterias, or services in boarding schools. The general case is that they are implicitly included, as few separate units exist. New Zealand reported some exceptions: a small number of boarding hostels attached to schools and universities. The United States indicated that they were included in the General Government sector but excluded from the Non Profit Institution Serving Households. Only Japan indicated that these ancillary services were excluded. This aspect is currently not taken into account in the temporal calculations of education volumes.

Weights

2.71 All countries have used cost weights, except Latvia, which considers the number of teachers (probably as a proxy of costs) and New Zealand, which uses a value added method.

Examples of methods from a selection of countries

2.72 The following section presents several examples from OECD countries for the measurement of education output. They were chosen because they represent different approaches and are therefore of more general interest.

Combined use of real earnings and scores data in Scotland

2.73 Exploratory work by Murray (2007) for Scotland has used a variant of the human capital approach (see Section 2.2.4) to deal with volume measurement of education output. The Scottish proposal is shown here because it is the most "outcome-based" method among OECD countries, and relies on both examination data and future real earnings.

2.74 Three categories of students of the upper secondary education are distinguished by level of attainments, i.e., by the qualifications gained over these three years of schooling. Category 1 represents the highest levels of qualifications gained; category 2 represents the middle level of attainment; and category 3 represents attainment at the low end of the attainment spectrum. Over the period 1997-2004, the study states that has been a movement from the lowest level of attainment (Category 3) to the medium level of attainment (Category 2). At the same time, the proportion of pupils gaining the highest level of attainment, Category 1, has remained fairly stable at around 30 per cent. This stratification by educational attainment captures the use of information about academic results, in the sense that education output is measured by the number of degrees obtained.

2.75 The human capital perspective enters via the weighting procedure: the weights produced were based on relative differences between the expected future earnings for the 3 categories of attainment. More specifically, as shown in the table below, average earnings for every category of attainment were adjusted by the employment rate and the expressed as a share of total earnings. The resulting percentages were averaged over the observation period and then used as weights for the changes in the number of students with different levels of attainment.

2.76 "This output index tracks the change in growth which can be caused by changes in the number of pupils attending school and changes in the proportion of pupils who fall into each attainment category. If these proportions remain constant throughout the period in question, there will be no change in quality and therefore output growth will solely be driven by changes in pupil attendance. If the proportions in each attainment category vary from year to year, then this will represent a change in the quality of education and will therefore influence overall output growth." (Murray 2007).

2.77 The results for the quality-adjusted output for the final year of pupils in secondary education is then projected to the other 11 years of education so as to gauge the total output from the school education system. A few comments are in place concerning the method. The first point is that weighting as suggested in the study implies that weights are reflective of *total real returns* following different levels of education. In principle, only the discounted value of the *incremental income* due to an extra level of education should be considered with potentially different results. This may be difficult empirically but would seem to be conceptually preferable.

2.78 The same idea of an incremental view should help answering the question about including those students in category 1 who follow secondary education with tertiary education and as a consequence fetch higher real earnings. Extra future real earnings due to tertiary education should be attributed to the tertiary education, and not the upper secondary education level. The problem is if all students of category 1 follow tertiary education, their future real earnings "without tertiary education" could be unobservable. In this case, a model would be necessary to attribute the respective parts to secondary and to tertiary education.

Table 2.5. Expected Weekly Wage in Real Terms by Highest Qualification Gained

	1999	2000	2001	2002	2003	2004	2005	Average Weight
Category 1								
Average earnings	£275	£283	£304	£298	£290	£297	£308	
Employment rate	0.75	0.77	0.78	0.76	0.77	0.77	0.79	
Expected earnings	**£206**	**£218**	**£238**	**£226**	**£223**	**£229**	**£242**	**48**
Category 2								
Average earnings	£211	£209	£230	£202	£229	£206	£212	
Employment rate	0.72	0.70	0.70	0.71	0.72	0.72	0.70	
Expected earnings	**£151**	**£145**	**£162**	**£143**	**£164**	**£149**	**£149**	**32**
Category 3								
Average earnings	£176	£196	£203	£201	£188	£209	£210	
Employment rate	0.47	0.48	0.49	0.50	0.50	0.51	0.50	
Expected earnings	**£83**	**£93**	**£99**	**£100**	**£94**	**£107**	**£104**	**20**

Source: Murray (2007)

Table 2.6. Quality-Adjusted Output Index for Final Year of Secondary Education

Number of Pupils	1997	1998	1999	2000	2001	2002	2003	2004
Category 1	18,189	17,899	18,769	18,486	18,583	18,479	18,494	18,315
Category 2	9,557	8,920	9,619	10,191	11,005	11,100	11,414	11,666
Category 3	31,014	29,358	28,388	28,203	28,024	28,079	28,870	28,531
Total number of pupils in cohort	61,659	59,071	58,652	59,250	60,138	59,998	61,035	60,447
Weighted total	1799210	1731725	1776452	1777500	1804621	1803780	1830318	1823082
Output index	**99.7**	**96.0**	**98.4**	**98.5**	**100.0**	**100.0**	**101.4**	**101.0**

Source: Murray (2007)

Finland: the use of ECTS credits in tertiary education

2.79 Heikkinen and Hautakangas (2007) report that Finland relies on a detailed stratification for education services, and distinguishes fields of education (although not exactly following ISCED), cross-classified by university and other education. There is also a split between "normal" and "special" classes in pre-primary, primary and lower secondary education. In upper secondary education, there is a distinction between general and vocational curricula.

2.80 Two institutional sectors and two information systems are solicited for measuring non-market education output: joint municipal boards (S1313) in pre-school, primary, lower secondary, vocational and upper secondary education, polytechnic activity, voluntary cultural activities of adult education centres and similar, basic art education and other educational services (26 indicators); and state sector (S1311) for universities and some vocational and general education, as well as further education of teachers (34 indicators). The KOTA database provides most of the data for central government. The activities of universities account for 95 per cent of central government's educational service activity.

2.81 Then, the best indicators for measuring volume output of education services at each level of education had to be identified. Different alternatives for volume indicators were studied and discussed with experts from the Board of Education and from Statistics Finland's education statistics. As student-hours were not available, and as the number of degrees was rejected because of the time-lag, the quantity indicators chosen comprised number of students, teaching-hours, number of credits and student-years, depending on the stratum.

2.82 The number of students was considered the best established indicator, and even preferred theoretically to student-hours in tertiary education. The number of pupils for a calendar year is obtained from reported pupil numbers by weighting data on the number of pupils in autumn of the statistical reference year and in autumn of the previous year with a 50% weight. The data on pupil numbers becomes available for the calculations approximately one month after the end of the statistical reference year.

2.83 Besides the number of students, credits are also used for polytechnics and universities. A problem with the use of student numbers was that students progress at varying speed and some of them do not earn a single credit during an academic year, and many of the latter do not study at all. The number of credits was then considered as better reflecting the educational output and outcome performed by the "true" students. Data on credits for the whole calendar year become available approximately 17 months from the end of the statistical reference year. Data on the spring semester, which become available about five months after the end of the statistical reference year, are used as preliminary data for the latest statistical reference year. Using the number of credits is recommendable practice, in particular for tertiary education, even though there may still be room for explicit quality adjustment. Student-years are used for the voluntary further vocational training only.

Table 2.7. Changes in education services by local governments in Finland

	Weight in 2003	2001	2002	2003	2004
Pre-school education (within the education system)	**0,016**	**1,246**	**1,068**	**0,989**	**1,012**
"Ordinary"	0,674	1,359	1,070	0,995	0,999
Severely disabled	0,070	1,078	1,084	1,023	1,035
Other disabled	0,256	1,011	1,058	0,968	1,040
Comprehensive education	**0,734**	**0,997**	**1,006**	**1,005**	**1,004**
"Ordinary"	0,900	0,987	0,996	0,996	0,990
Severely disabled	0,007	0,997	1,046	1,057	1,021
Other disabled	0,030	1,083	1,074	1,065	1,064
Special education	0,063	1,155	1,146	1,112	1,096
Upper secondary general education	**0,127**	**0,990**	**0,975**	**0,974**	**0,979**
Under 18-year olds	0,031	0,930	0,926	0,945	0,958
Over 18-year-olds	0,969	0,993	0,978	0,976	0,980
Vocational education	**0,050**	**1,028**	**1,077**	**0,995**	**1,013**
Natural resources	0,085	1,020	0,904	1,021	1,063
Technology, communications and transport	0,393	1,021	1,047	1,018	1,002
Social sciences, business and administration	0,096	1,062	1,035	1,009	0,993
Tourism, catering and domestic services	0,125	0,996	1,006	1,010	0,988
Social services and health	0,130	1,028	1,052	1,016	1,025
Culture	0,081	1,123	1,081	1,079	1,060
Recreation and sports	0,004	1,538	1,176	1,129	1,180
Apprenticeship training	0,045	0,961	1,008	1,019	0,964
Further education	0,040		4,644	0,602	1,047
Polytechnic education	**0,048**	**1,078**	**1,003**	**1,010**	**1,003**
Natural resources	0,049	1,127	1,181	0,972	
Technology, communications and transport	0,398	1,117	1,063	0,971	
Social sciences, business and administration	0,165	1,027	1,034	0,985	
Tourism, catering and domestic services	0,038	1,332	1,063	1,073	
Social services, health and sports	0,201	1,037	1,010	1,028	
Culture	0,141	1,215	1,229	1,121	
Humanities and education	0,007	1,484	1,153	1,035	
Adult education centres	**0,026**	**1,007**	**1,007**	**0,998**	**1,020**
VOLUME INDICATOR	**1,00**	**1,003**	**1,006**	**1,000**	**1,002**
Change in volume (old method)		1,012	1,022	1,008	1,019
Difference, volume indicator – old method		-0,009	-0,016	-0,008	-0,017

Source: Heikkinen and Hautakangas (2007).

2.84 Finland privileges an approach towards output measurement by economic unit over an approach by product, research conducted by universities is part of "educational output", and has been estimated by the number of publications. This quantity indicator is convenient, although research expenditure should be treated as a separate product, along with other R&D according to the OECD *Frascati Manual*.

Table 2.8. Changes in education services by central governments in Finland

	Weight in 2003	2001	2002	2003	2004
General education	**0.076**	**0.978**	**1.010**	**1.000**	**1.003**
Pre-school education	0.031	0.962	0.947	0.948	0.946
"Ordinary"	0.481	0.962	0.993	1.050	1.028
Disabled education	0.519	1.000	0.922	0.859	0.869
Comprehensive education	0.791	0.975	1.010	1.002	1.003
"Ordinary"	0.640	0.994	0.993	1.004	1.011
Disabled education	0.340	0.937	1.051	0.988	0.988
Special education	0.020	0.924	0.957	1.167	1.028
Upper secondary general education	0.179	1.003	1.023	1.000	1.009
Vocational education	**0.021**	**1.012**	**1.048**	**1.088**	**1.015**
Natural resources	0.113	1.057	1.066	1.010	1.056
Technology, communications and transport	0.269	1.000	1.031	1.164	1.036
Social sciences, business and administration	0.102	1.329	1.079	1.000	0.941
Tourism, catering and domestic services	0.250	1.000	1.078	1.109	1.010
Social services and health	0.062	0.611	0.777	1.333	1.042
Culture	0.204	1.044	1.098	1.010	1.000
University education	**0.904**	**1.024**	**1.045**	**1.031**	**1.031**
Basic degree programmes	0.383	1.038	1.048	1.027	1.021
Theology	0.009	1.061	1.042	1.114	1.043
Humanities	0.111	1.018	1.015	1.028	1.035
Art and design	0.043	1.056	1.041	1.003	0.942
Music	0.030	0.998	1.051	0.985	0.938
Theatre and dance	0.016	1.055	1.141	1.056	0.938
Education	0.107	1.015	1.017	1.003	1.006
Sport sciences	0.006	1.070	1.111	1.027	1.032
Social sciences	0.068	1.035	1.032	1.020	1.034
Psychology	0.009	1.024	0.995	1.082	0.951
Health sciences	0.011	1.064	1.027	1.076	1.025
Law	0.017	0.990	0.998	1.043	1.021
Economics	0.081	1.057	1.061	1.039	1.024
Natural sciences	0.154	1.058	1.063	1.024	1.025
Agriculture and forestry	0.029	1.005	1.014	1.002	1.031
Engineering	0.200	1.047	1.070	1.020	1.025
Medicine	0.074	1.038	1.062	1.095	1.084
Dentistry	0.008	1.073	1.076	1.080	1.102
Veterinary medicine	0.011	0.995	1.073	1.008	0.908
Pharmacy	0.011	1.076	1.035	0.932	0.951
Fine arts	0.005	1.077	0.955	1.028	1.338
Higher degree programmes	0.252	1.023	1.044	1.047	1.018
Research	0.364	1.011	1.042	1.023	1.089
VOLUME INDICATOR	**1.00**	**1.020**	**1.042**	**1.029**	**1.041**
Change in volume (old method)		1.053	1.043	1.015	0.989
Difference, volume indicator – old method		-0.033	-0.001	0.014	0.052

Number of students
Number of credits
students
publications

Source: Heikkinen and Hautakangas (2007).

Italy: the use of a class-size model and the accounting of actual time for graduation in tertiary education

2.85 Collesi, Guerrucci, Versace and Zannoni (2007) describe the Italian method of measuring education services. Italy has followed the Eurostat handbook on prices and volume in national accounts. As the teaching hours per students are not available in Italy, the Italian quantity indicator is the number of enrolled students. In addition, indicators have been developed to reflect the quality of the service provided in the two major activities. The Italian education output is divided into four activities, of which the school system and the university education are the major ones.

Table 2.9. Composition of education output by type of service in Italy Percentage shares at current prices

Type of service	2000	2001	2002	2003	2004	2005
School system	87.2	86.7	86.5	86.8	85.9	86.2
Vocational training	4.3	4.5	4.3	4.1	4.5	4.4
University education	8.1	8.3	8.7	8.6	9.0	8.8
Subsidiary services to education	0.4	0.5	0.6	0.6	0.5	0.5
Total	100.0	100.0	100.0	100.0	100.0	100.0

Source: Collesi, Guerrucci, Versace and Zannoni (2007)

2.86 The education part of the school system is divided into four levels: pre-primary education, primary education, lower secondary and upper secondary education (ISCED-97 levels 0-3). Education is supplied mainly in state schools, but other public schools can be managed by local authorities: Municipalities, Provinces and Regions. The volume index is calculated at the lowest level of aggregation. This means that the number of students in state and non-state schools (quantity index) is broken down into the four levels of education and, in upper secondary education, by type of institute (classical lyceum, scientific lyceum, teacher training institutes and schools, vocational institutes, technical institutes, art institutes, art lyceums). This last sub-stratification of upper secondary education is by field of education.

2.87 The quality adjustment coefficient is based on the number of pupils per class. Classes are divided by the level of education as well. The idea is that if the number of students per class increases the individual attention that a teacher may dedicate to each of them decreases. Based on these observations, a conversion function has been constructed, which takes on a linear form for the indicator values below the level at which congestion starts (q = n), and a non-linear form, similar to a conventional production function, for higher values. Furthermore, the function reaches its maximum in the point of maximum capacity, beyond which the service is overcrowded. The conversion function transforms the actual number of pupils into a number of "real" or normal-equivalent pupils who have received a service that can be defined as “standard” quality.

(1) $$q = f(n) = \begin{Bmatrix} n \\ -1/30n^2 + 5/3n - 10/3 \end{Bmatrix} \text{for} \begin{Bmatrix} 0 \le n \le 10 \\ n > 10 \end{Bmatrix}; \; \max q = f(25)$$

2.88 For instance, 25 pupils per class (n) allow the teacher to dedicate attention to an equivalent of 17.5 ‘normal’ equivalent pupils (q), which is the maximum reached by this function. The Italian model and the corresponding chart have already been exposed in this handbook. They constitute a useful way of dealing with quality adjustment.

2.89 Activity of universities in Italy has been split in two CPA classes: R&D, deflated separately by an input method, and education *stricto sensu*, stratified by 18 groups of faculties, reflecting the fields of education. A model is used to determine the unit cost per student, based on the number of students and on the number of equivalent professors.

2.90 Italy has chosen two outcome-based indicators:

- The ratio between the enrolled "regular students" in the course SC_{jt} (i.e. students who did not exceed the legal length of their degree) and the total number of enrolled students S_{jt};

- The reduction of the distance between the actual number of years for graduation LE_{jt} and the theoretical length LT_{jt}.

2.91 The correction factor applied is: q_{jt}/q_{jt-1} where $q_{jt} = ((SC_{jt}/S_{jt}) + (LT_{jt}/LE_{jt}))/2$ for each group of faculties j and each period t. The first indicator, close to 1 for "efficient" faculties, reflects the classical "outcome-based" idea of "pupils moving up" rather than "enrolled students".

2.92 The second idea is more adapted to the precise situation of Italian faculties after 1999: European harmonisation of degrees led to a re-organization of all curricula, and, on average, to a reduction of the length of graduation time. It is assumed that the "value" of the final degree has remained constant. At the end of the reform, the ratio should be close to 1, too.

Table 2.10. Actual and theoretical average time for degrees in Italy

Groups of faculties	2000			2004		
	LE	LT	LT / LE	LE	LT	LT / LE
01 Ssciences	10.18	4.65	0.46	6.90	3.91	0.57
02 Pharmacy	11.07	4.98	0.45	7.94	4.75	0.60
03 Medicine and surgery	7.52	5.88	0.78	5.88	3.79	0.64
04 Engineering	12.78	5.00	0.39	6.97	4.05	0.58
05 Architecture	15.05	4.98	0.33	8.02	4.39	0.55
06 Agriculture	11.60	5.00	0.43	7.38	4.20	0.57
07 Veterinary medicine	12.20	5.00	0.41	8.59	4.75	0.55
08 Sociology	7.68	4.15	0.54	7.89	3.80	0.48
09 Political science	8.67	4.00	0.46	6.64	3.61	0.54
10 Law	8.65	4.00	0.46	8.24	3.81	0.46
11 Letters	8.47	4.04	0.48	7.16	3.76	0.53
12 Language	8.46	4.00	0.47	6.88	3.72	0.54
13 Cultural heritage	8.46	4.00	0.47	8.24	3.83	0.46
14 Psychology	10.03	5.00	0.50	6.43	4.24	0.66
15 Economics	8.76	4.00	0.46	6.75	3.63	0.54
16 Education	7.40	4.09	0.55	6.78	3.82	0.56
17 Statistics	7.84	4.00	0.51	6.21	3.45	0.56
18 Exercise and sport science	4.00	4.00	1.00	4.48	3.42	0.76
Total	9.75	4.43	0.45	7.00	3.87	0.55

Source: Collesi, Guerrucci, Versace and Zannoni (2007)

2.93 The subsidiary services to education concern all the activities that support the university studies cycle. The main activity, classified here in "education" in the sense of the forthcoming ISIC rev 4, is measured in volume by the number of meals supplied and by the number of bed places assigned and the prices are represented by the cost of production of the services. All secondary productions, which are allocated outside of the diagonal in the Supply table, are deflated with a suitable index in relation to the products originating from the secondary productions.

CHAPTER 3. MEASURING EDUCATION SERVICES ACROSS COUNTRIES

3.1 Introduction

3.1. In May 2007, the Eurostat Working Party on National Accounts, jointly with the Working Party on Purchasing Power Parities, discussed a report prepared by a Eurostat Taskforce on the measurement of education services across countries. The report recommended a method based on pupil and student numbers, differentiated by level of education as the basic quantity measure and, for primary and secondary level of education, a quality adjustment on the basis of PISA, an international student assessment carried out by OECD. The Task Force carried out test calculations and concluded that the method provides – on the whole – more plausible results than the current input method.

3.2. Indeed, there are significant qualitative differences in education systems across countries that cannot be captured by the input method. Moreover, output-based approaches are recommended and increasingly implemented in national accounts. As temporal and spatial comparisons should be consistent, this provides an added rationale for an output-based approach in spatial comparisons.

3.3. Measurement of PPPs for education services has traditionally been based on an input method. This meant that PPPs were developed for various input components and then aggregated to derive a PPP for education as a whole. The most important input component being wages and salaries, much of the PPPs for total education depended on the quality of wage and salary comparisons of teachers and other personnel across countries. Experience showed that such comparisons were difficult and tended to produce unreliable results. Moving towards an output method in this area is thus not only a conceptual improvement, it also bears the possibility of increasing data quality.

3.2 Temporal and spatial dimension - differences in measurement

3.4. Measurement of volume for education services across countries should be symmetric to the measurement of volume over time. This does not only concern the general principles but should also relate to aspects of implementation. The main common features are:

- As in PPP comparisons in general, volume measures of educational services – everything else being equal - should not be affected by differences in the shares of market and non-market activities. In spatial comparisons, the target measure for volumes and PPPs is *actual individual consumption* of educational services, i.e. the sum of expenditure of households, NPISHs and general government on education;

- In the absence of economically significant prices, comparisons across countries rely either on direct volume measures or on unit costs that are compared across countries and used to deflate education expenditure;

- There are plenty of data available to measure the quantity of educational services but capturing the quality of the services remains an issue. Due to wide differences between countries, the relative importance of quality is much higher in a spatial context than in a temporal context and finding data to carry out a quality adjustment remains challenging.

3.5. In PPP comparisons, like in temporal price and volume indices, products to be compared should be representative, comparable and consistent. These requirements are valid independently of whether the comparison is based on price data or whether volume data is taken as starting point as in the case of educational services - reliable results cannot be achieved without meeting these requirements.

- *Representativity* means that services to be compared represent adequately the whole expenditure category;

- *Comparability* means that the units used in the volume estimation (e.g. service received by a student in the comparison year) should be the same across countries. Moreover, in an indirect estimation of PPPs as a ratio between expenditure and volume, underlying expenditure and volume data should be consistent with each other. This ensures that resulting PPPs are based on comparable "prices" in the same way as in those parts of comparison where the PPP estimation relies on direct use of price data;

- *Consistency* means that the identity price x volume = expenditure should hold. In the framework where the volume data are exhaustive and the comparability requirement of PPPs is met, possible differences in the national accounts expenditure data across countries show up as volume differences rather than price differences.

3.6. There is often a trade-off between the representativity and comparability requirements. High requirements for representativity may result in lower comparability and vice versa. For educational services, quantitative data on the number of students etc. tend to be widely available, and while not perfect, they are appropriate to be used in the estimation. Thus, the representativity requirement is met and the main problem is how to ensure that the nature and quality of education services are comparable across countries.

3.7. The following steps describe the basic procedure to estimate output-based PPPs or their equivalent in form of a direct volume index:

- Stratification of expenditure on education services into homogeneous groups;

- For each stratum, identification of the quantity measure of education services;

- For each stratum, identification of the quality measure of education services. By combining it with the quantity indicator, a quality-adjusted volume or a spatial price index (PPP) can be derived.

3.3 Estimation of output-based PPPs for education services

3.3.1 Stratification of education services

3.8. A correct estimation procedure requires that educational services are stratified into homogeneous groups where ISCED levels, programme orientations and the quality of services are taken into account. At the international level, the classification available for the stratification is ISCED that also underlies the UOE (UNESCO, OECD, Eurostat) data collection system on education. In practice, possibilities for detailed stratification are very limited because of a lack of comparable base data. The Eurostat Taskforce on education PPPs ended up proposing a fairly aggregate stratification according to ISCED:

ISCED 0	Pre-primary education
ISCED 1	Primary education or first stage of basic education
ISCED 2	Lower secondary or second stage of basic education
ISCED 3+4	Upper secondary and post-secondary non-tertiary education
ISCED 5+6	Tertiary education (including category "unknown")

3.9. Whether this stratification is satisfactory, depends on the comparability of educational services within one stratum across countries. If services are very different across countries, differences in production costs will translate into price differences rather than volume differences and potentially bias comparisons. However, as all countries are likely to offer a wide set of educational services within each ISCED heading it is not obvious that such heterogeneity results in major errors, unless the average quality in one country's education services is significantly different from the quality in a comparison country.

3.3.2 Quantity of education services

3.10. For temporal comparisons, the basic quantity indicator for education has been identified as the number of "pupil-hours", i.e. the number of hours that pupils spend being taught. Alternatively, the number of pupils could be compared. The choice between the two measures is likely to be of greater importance in a spatial than in a temporal context, because the number of pupil-hours per pupil varies significantly more across countries than it varies across years within a country. Unless one believes that teaching services are strictly linked to the number of hours per pupil, a case could be made for comparing the number of pupils. One advantage when using the number of pupils is that comparative data for quality adjustment is based on pupils rather than on pupil-hours.

3.11. Furthermore, data on student-hours are in general available only for primary and secondary levels of education. Instead, information on numbers of students and on numbers of full-time equivalent (FTE) students is widely available in the UOE data collection system. Of these concepts, the number of FTE students can be considered to approximate services received and is closer to the concept number of student-hours. Therefore, it has been taken as a basis for the estimation of quantity of education services.

3.12. It should be noted that the definition of FTE student is not necessarily uniformly applied across countries. Some OECD countries count every participant at the tertiary level as a full-time student while others determine a student's intensity of participation by the credits which he or she obtains for successful completion of specific course units during a specified reference period. The influence of differences in definition has not yet been investigated.

3.13. In PPP comparisons, the main objective is to measure volumes underlying total actual individual consumption of education. Thus the targeted quantity measure relates to all students in all types of educational institutions (public and private).

3.14. An important issue is how to reconcile differences between expenditure data on education as available from the UOE data collection and expenditures on education in national accounts. The results are very sensitive to the choice of the method. Any automatic adjustment of expenditure data is risky because reasons for inconsistencies in data may differ between countries. More work will be needed to identify sources of differences in expenditure data and the best way to reconcile sources for each country.

3.3.3 Quality of education services

Alternative quality adjustment methods

3.15. The discussion on measuring education services over time already showed that a method based on unadjusted quantities for education services would not be fully satisfactory, and some quality adjustment may be needed. This conclusion holds even more in a spatial context because it is likely that quality differences across countries are relatively more important than quality differences over time within a country. Some quality adjustment is thus called for. It is clear, however, that quality of education is a highly sensitive issue and its measurement is fraught with conceptual and practical difficulties.

3.16. As discussed in the previous chapter, there are several approaches to adjust directly education services. Eurostat (2001) lists the following (section 3.1.2.2, page 34):

- Direct measurement of the quality of the output itself. In the case of education, reports of school inspections could be used. However, this does not lend itself for use in inter-country comparisons;
- Measuring the quality of the inputs. Using teacher/student ratios or class size information is an example of such an approach. The Handbook says: "An assumption is then made that the quality change of the inputs leads automatically to a quality change of the output. However, this assumption cannot be verified without actually measuring the quality of the output." Results would not be satisfactory;
- Using outcomes. The Eurostat Handbook says: "The quality of the output lies in its results, i.e. in the outcome. The most appropriate way of adjusting for quality is to investigate changes in outcome indicators". For education, this implies investigating for example examination results.

3.17. In the event, and for practical reasons, only an adjustment based on the international studies of the attainment levels of students was considered[16]. Even there possibilities are limited because adequate data are not in general available. Only a few studies exist that concern education services at the primary and secondary level:

> **PISA**: at the moment, PISA seems to be the most reliable and complete international database for secondary schools. The Programme for International Student Assessment is an internationally standardised assessment managed by OECD, which was jointly developed by participating countries and administered to 15-year-olds in schools. This three-yearly survey was implemented in 43 countries in the first assessment in 2000, in 41 countries in the second assessment in 2003 and 57 countries participated in the third assessment in 2006. 62 countries have signed up to participate in the 4th assessment in 2009. Tests are typically administered to between 4 500 and 10 000 students in each country. There are three topics of examination in PISA: mathematics, reading and science. In each round, there is particular emphasis on one of the three topics.

16 Other possibilities would have included school completion rates, transition from secondary to tertiary education, successful transition of students with disabilities into independent living arrangements etc.

PIRLS: the project Progress in International Reading Literacy Study gives qualitative indicators for primary schools. 35 Countries participated in PIRLS 2001, testing 150.000 9-and-10-year-olds in schools. A new PIRLS survey was carried out in 2006 where the number of participating countries rose to 40. It is managed by the International Study Centre (ISC) at Boston College, in partnership with the International Association for the Evaluation of Educational Achievement (IEA).

TIMSS: The Trends in International Mathematics and Science Study has carried out in 1995, 1999, 2003 and in 2007. It is undertaken by 9-year-olds and 13-year-olds, involving 60 countries in total but rather few countries of the EU. It is also managed by ISC and IEA.

3.18. From these three surveys, PISA has the widest coverage of education fields and countries, and seems the most secure in terms of future continuation. Another advantage of PISA is that it provides results that are corrected for the economic, social and cultural status of students (the so-called ESCS-correction), thus making possible to evaluate better the quality of teaching.

Adjustment of primary and secondary education services based on PISA results

3.19. The basic approach of PISA is to measure the level of attainment of 15-year-olds by testing them in the three fields of knowledge mentioned above. The tests are the same in each country. The scores are subsequently placed on a scale with 500 as the average of OECD countries and 100 as standard deviation.

3.20. PISA scores can be located along specific scales developed for each subject area, designed to show the general competencies tested by PISA. These scales are divided into levels that represent groups of PISA test questions, beginning at Level 1 with questions that require only the most basic skills to complete and increasing in difficulty with each level. Once a student's test has been corrected, his or her score in reading, mathematics and science (plus problem solving in PISA 2003) can be located on the appropriate scale. For example, a student who is likely to lack the skills needed to correctly complete the easiest questions on a PISA test would be classified as below Level 1, while a student who is likely to have many of the skills needed to correctly complete the test questions would be at a higher level. In each test subject, the score for each participating country is the average of all student scores in that country. The average score among OECD countries is 500 points and the standard deviation is 100 points. About two-thirds of students across OECD countries score between 400 and 600 points.

3.21. There are a number of problems to be solved when using score data for quality adjustment. These were already discussed in chapter 2 but might my useful to briefly recall in the present context because the magnitude of problems tends to be bigger in spatial comparisons than in national accounts due to wide differences between countries. The main problems are the following:

- How to ensure that score data used reflect the quantity and quality of teaching rather than pupils' personal ability and socio-economic factors?
- How to ensure that score data used in adjustment are equally representative across countries in spite of possible differences in the curriculum?
- Perhaps the most critical question: how to translate scores into a metric scale?

3.22. To eliminate socio-economic and environmental factors, PISA scores that have been adjusted for economic, social and cultural status of pupils ('ESCS correction') can be used for quality adjustment of student numbers. In the test calculations, score data represent un-weighted arithmetic averages of scores for three tests – mathematics, sciences and reading. These three tests can be understood to cover a significant share of teaching in schools and it seems that the share is not very different across countries. It is thus unlikely that a separate treatment of the three score results would change results significantly. Influence of a bigger share of e.g. mathematics in the curriculum may improve math scores for that particular country but this may be compensated by lower scores in the other tests. An open question is how well the fields covered by tests represent the whole curriculum.

3.23. As explained above, PISA score data have been normalised to an average score of 500 and a standard deviation of 100 based on all OECD countries. The choice of average and standard deviation is arbitrary, and both influence directly the distance between countries. A higher score average and/or lower standard deviation would decrease the relative difference between countries and vice versa.

3.24. The test calculations (see Table 3.2) show the effects of quality-adjustment with score data. The general impression is that of a moderate influence and about 8 per cent at maximum.

3.25. By way of conclusion on spatial comparisons of education services, the number of full-time equivalent students differentiated by level of education is a reasonable way forward, and has been put in place in the Eurostat PPP programme. Possible differences in the conversions to full-time equivalents merit further investigation. An explicit quality adjustment is appropriate given the likely differences in educational quality between countries. PISA results, adjusted for socio-economic factors proved to be feasible for primary and secondary education as shown by test calculations. No quality adjustment has so far been put in place for tertiary education services and development work is required in this area.

ANNEX 3.A: RESULTS OF EXPERIMENTAL CALCULATIONS FOR 2005

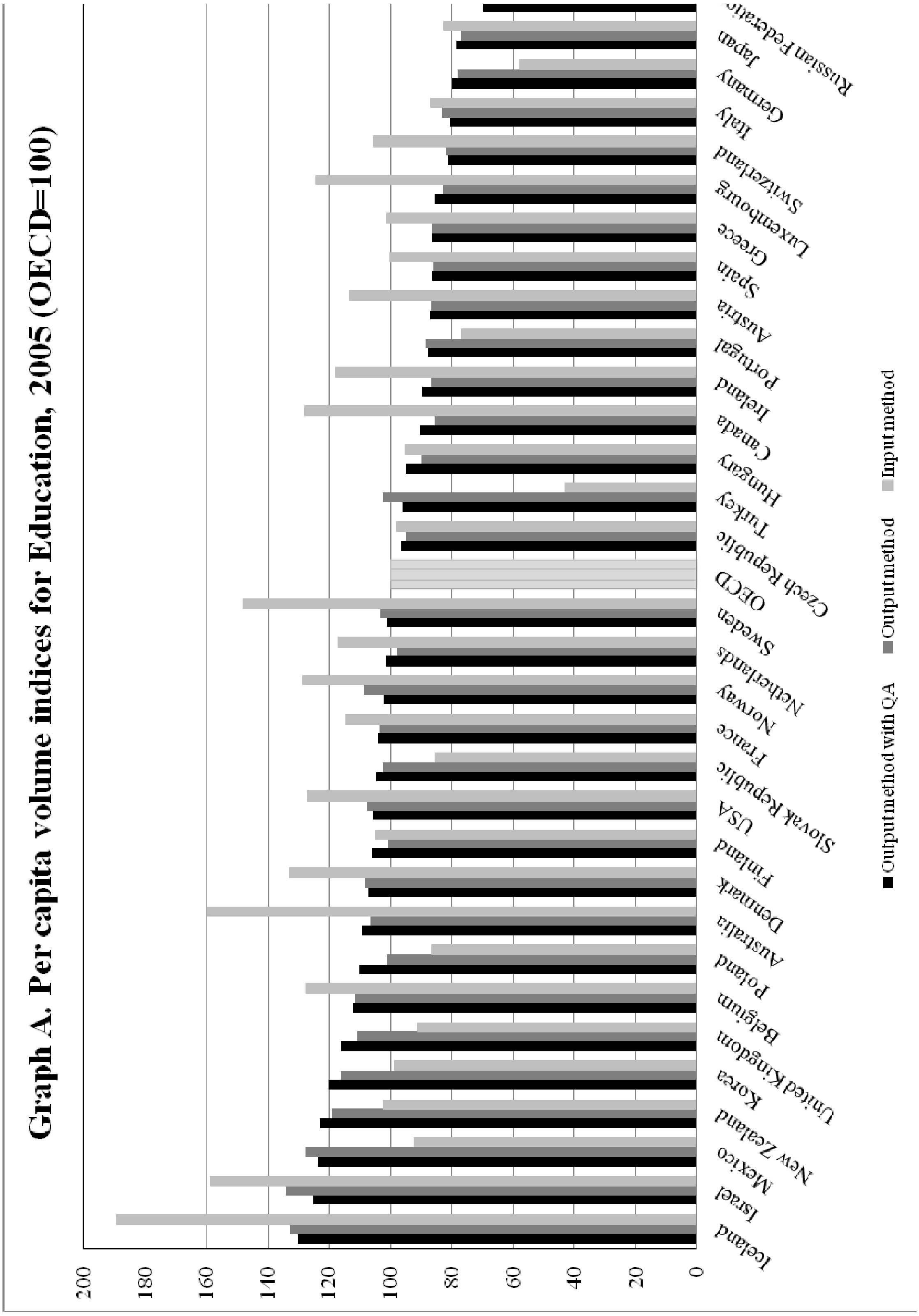
3.26. Experimental calculations have been carried out for the year 2005 covering the 30 OECD countries plus Israel and the Russian Federation. Tables and graphs showing the main results are presented below. Due to the experimental character of the calculations, results should be read with caution. 2005 base data were incomplete for several countries and were estimated based on previous years' data. Nevertheless, the results show the direction and magnitude of changes that can be expected when an output approach is adopted.

3.27. Table 3.1 and the following figure show that variations in results are much stronger in the case of the input approach than in the case of the output approach. In the input method, the index ranges from 43 (Turkey) to 189 (Iceland) whereas it varies between 70 (Russian Federation) and 130 (Iceland) under the output approach. For countries such as Iceland, Australia and Sweden, extremely high volumes in the input approach reduce to a more plausible level when an output method is applied. The introduction of the quality adjustment with PISA seems to be rather limited influence on results. The adjustment is biggest for Poland where the index goes up by 8% (from 101 to 110). Table 3.2 shows the impact of methods on the level of GDP per capita and on the level of actual individual consumption. The relatively small effect reflects the share of education services in GDP of about 5 per cent and OECD countries.

Table 3.1. Indices of real final expenditure per head on education (OECD=100)

	Output method				Input method	rank	% change between input and output QA
	with QA	rank	without QA	rank			
Iceland	130	(1)	133	(2)	189	(1)	-45.5
Israel	125	(2)	134	(1)	159	(3)	-27.0
Mexico	124	(3)	128	(3)	92	(24)	25.4
New Zealand	123	(4)	119	(4)	103	(17)	16.4
Korea	120	(5)	116	(5)	99	(21)	17.7
United Kingdom	116	(6)	111	(7)	91	(25)	21.6
Belgium	112	(7)	112	(6)	128	(8)	-13.7
Poland	110	(8)	101	(16)	87	(27)	21.2
Australia	109	(9)	106	(11)	159	(2)	-45.8
Denmark	107	(10)	108	(9)	133	(5)	-24.1
Finland	106	(11)	101	(17)	105	(16)	1.1
USA	106	(12)	108	(10)	127	(9)	-20.4
Slovak Republic	105	(13)	103	(15)	85	(28)	18.6
France	104	(14)	104	(12)	115	(13)	-10.2
Norway	102	(15)	109	(8)	129	(6)	-26.3
Netherlands	102	(16)	98	(19)	117	(12)	-15.4
Sweden	101	(17)	103	(13)	148	(4)	-46.7
OECD	100	(18)	100	(18)	100	(20)	0.0
Czech Republic	97	(19)	95	(20)	98	(22)	-1.3
Turkey	96	(20)	103	(14)	43	(33)	55.3
Hungary	95	(21)	90	(21)	95	(23)	-0.1
Canada	91	(22)	85	(27)	128	(7)	-41.6
Ireland	90	(23)	87	(23)	118	(11)	-31.5
Portugal	88	(24)	88	(22)	77	(31)	12.5
Austria	87	(25)	87	(24)	114	(14)	-30.3
Spain	87	(26)	86	(26)	100	(19)	-15.9
Greece	86	(27)	86	(25)	101	(18)	-17.0
Luxembourg	86	(28)	83	(29)	124	(10)	-45.5
Switzerland	81	(29)	82	(30)	106	(15)	-30.0
Italy	81	(30)	83	(28)	87	(26)	-8.1
Germany	80	(31)	78	(31)	58	(32)	27.4
Japan	79	(32)	77	(32)	83	(29)	-5.1
Russian Federation	70	(33)	69	(33)	80	(30)	-14.1

* QA: Quality adjustment



Graph A. Per capita volume indices for Education, 2005 (OECD=100)

Table 3.2. Indices of real final expenditure per head on GDP at average OECD prices (OECD=100)

	Input method 2005			Quantity approach			Quantity approach + quality adjustment (PISA + ESCS)		
	Education	AIC	GDP	Education	AIC	GDP	Education	AIC	GDP
Austria	114	111	117	87	109	116	87	109	116
Belgium	128	103	111	112	102	110	112	102	110
Czech Republic	98	63	70	95	63	70	97	63	70
Denmark	133	102	116	108	101	115	107	101	115
Finland	105	93	105	101	93	105	106	93	105
France	115	106	102	104	105	102	104	105	102
Germany	58	103	105	78	106	107	80	106	107
Greece	101	88	88	86	87	87	86	87	87
Hungary	95	59	59	90	58	59	95	59	59
Iceland	189	127	127	133	124	125	130	124	125
Ireland	118	100	131	87	98	129	90	98	129
Italy	87	93	95	83	93	95	81	93	95
Luxembourg	124	159	239	83	154	235	86	155	235
Netherlands	117	107	119	98	106	119	102	106	119
Norway	129	117	162	109	116	162	102	116	161
Poland	87	49	47	101	50	47	110	50	48
Portugal	77	73	69	88	74	70	88	74	70
Slovak Republic	85	53	55	103	53	55	105	53	55
Spain	100	91	94	86	90	93	87	90	93
Sweden	148	104	110	103	101	108	101	101	108
Switzerland	106	111	123	82	109	122	81	109	121
Turkey	43	27	27	103	29	28	96	29	28
United Kingdom	91	119	109	111	121	110	116	121	111
Australia	159	104	113	106	101	111	109	101	111
New Zealand	103	85	85	119	86	86	123	86	86
Japan	83	97	104	77	97	104	79	97	104
Korea	99	58	74	116	59	74	120	59	75
Canada	128	112	121	85	109	119	91	109	119
Mexico	92	42	39	128	43	40	124	43	40
USA	127	152	143	108	150	143	106	150	142
Israel	159	72	78	134	72	77	125	71	77
Russian Federation	80	37	40	69	37	40	70	37	40
OECD	100	100	100	100	100	100	100	100	100

Table 3.3. Number of students, 2005

	Number of students	Population total (thousands)	Students as percentage of total population	Volume of education services per capita, OECD=100
Austria	1678	8236	20.38%	91
Belgium	2678	10479	25.55%	114
Czech Republic	2158	10236	21.09%	94
Denmark	1353	5419	24.96%	112
Finland	1295	5246	24.69%	110
France	14926	62818	23.76%	106
Germany	16617	82469	20.15%	90
Greece	2181	11104	19.64%	88
Hungary	2152	10087	21.33%	95
Iceland	90	297	30.23%	135
Ireland	1013	4159	24.36%	109
Italy	11035	58607	18.83%	84
Luxembourg	75	465	16.12%	72
Netherlands	3580	16320	21.94%	98
Norway	1155	4623	24.97%	112
Poland	9068	38165	23.76%	106
Portugal	2175	10549	20.62%	92
Slovak Republic	1254	5387	23.27%	104
Spain	8520	43398	19.63%	88
Sweden	2226	9030	24.65%	110
Switzerland	1459	7437	19.62%	88
Turkey	16394	72065	22.75%	102
United Kingdom	15015	60227	24.93%	112
Australia	4544	20340	22.34%	100
New Zealand	1075	4099	26.22%	117
Japan	21031	127768	16.46%	74
Korea	11608	48138	24.11%	108
Canada	6068	32299	18.79%	84
Mexico	31748	103947	30.54%	137
United States	67267	296507	22.69%	102
OECD	261437	1169922	22.35%	100
Israel	2066	6930	29.82%	133
Russian Federation	21309	143114	14.89%	67

Table 3.4. PISA scores, 2006

	PISA scores				Adjustment factor 2006
	Math.	Sciences	Reading	Average	
Austria	505	511	490	502	*1.003*
Belgium	520	510	501	510	*1.017*
Czech Republic	510	513	483	502	*1.006*
Denmark	513	496	494	501	*0.986*
Finland	548	563	547	553	*1.098*
France	496	495	488	493	*1.003*
Germany	504	516	495	505	*1.000*
Greece	459	473	460	464	*0.944*
Hungary	491	504	482	492	*0.996*
Iceland	506	491	484	494	*0.944*
Ireland	501	508	517	509	*1.027*
Italy	462	475	469	469	*0.952*
Luxembourg	490	486	479	485	*0.972*
Netherlands	531	525	507	521	*1.038*
Norway	490	487	484	487	*0.935*
Poland	495	498	508	500	*1.030*
Portugal	466	474	472	471	*0.989*
Slovak Republic	492	488	466	482	*0.977*
Spain	480	488	461	476	*0.982*
Sweden	502	503	507	504	*0.989*
Switzerland	530	512	499	514	*1.031*
Turkey	424	424	447	432	*0.925*
United Kingdom	495	515	495	502	*1.010*
Australia	520	527	513	520	*1.023*
New Zealand	522	530	521	524	*1.032*
Japan	523	531	498	517	*1.065*
Korea	547	522	556	542	*1.100*
Canada	527	534	527	529	*1.025*
Mexico	406	410	410	409	*0.926*
USA	474	489	..	*486*	*0.952*
Israel	..	454	439	*447*	*0.891*
Russian Federation	476	479	440	465	*0.930*

Figures in italics are OECD estimates.

CHAPTER 4. MEASURING HEALTH SERVICES OVER TIME

4.1 Terminology and concepts in health

4.1.1 Introduction

4.1. This chapter of the Handbook provides guidance on the development of an output-based measure of the volume of health services to be used for national accounting purposes and, more generally, for comparing volumes of health expenditures.

4.2. Accurate and comparable measures of both the size and the growth of the health sector are increasingly important for a number of reasons:

- Health spending is increasing across OECD countries and is accounting for a growing share of GDP;
- Governments and citizens are interested in knowing whether health funds are well spent for purposes of accountability and resource allocation, and for informing individual choices;
- International comparisons are one of the most powerful mechanisms for evaluation of and change in national health systems. International comparisons require good quality data which is consistent and sufficiently representative for countries;
- As the ultimate goal of health services is to improve people's health, there is a growing interest in the quality of health services. In the case of measuring volume output of health, a health activity with a higher composite quality than another health activity could be identified as such if it contributes more to health outcomes than the alternative activity.

4.3. The challenges of measuring the output of health providers arise from two key elements of the nature of health care (Office of Health Economics, 2008). First, it is largely a customised rather than a standardised product, with a complex production process. Second, there is an absence of market or economically significant prices for many services and goods – they are frequently provided by non-market producers. The standard method of measuring the value of output has been by summing costs. Changes in the volume (or price) of health services have typically been captured by measuring the changes in the volume (or price) of inputs.

4.4. However, health volume output should be measured as the quantity of health services provided to individuals with an adjustment for new products or services and quality change and not as the quantity of inputs used to produce these services. In the latter case, the measures of real output in an economy are incomplete as are measures of consumption and real income. Also, in the absence of volume measures of output that are independent from volume measures of inputs, no meaningful productivity measures can be constructed.

4.5. As in the case of education, capturing quality change in the provision of health services constitutes a challenge for statisticians and analysts and some space will be allocated to the discussion of this issue in the present chapter. One of the features of the health industry distinguishes that it from formal education is the wide and heterogeneous range of activities in health care. There are several thousands of diseases in the International Classification of Diseases (ICD), and the additional complexity of co-morbidities. Associated with the thousands of diseases are a very large number of interventions/treatments for these different diseases. Furthermore, there is variation in the mix of activities which may be applied for the same type of intervention.

4.6. In consistency with the general principles of volume output measurement set out earlier in this Handbook, the quantity and quality indicators that are recommended for use in a time series analysis will be, at least conceptually, the same as the quantity and quality indicators to be used in cross-country comparisons. For example, if it is proposed to use identical treatments to measure the growth of the volume of hospital services *over time* and for a given country, symmetrically, for comparisons of output of hospital services *of two countries* at a given point in time, it will also be recommend using an aggregate based on comparable treatments in the two countries.

4.7. The chapter starts out with a definition of health care, and then goes on to discuss and define inputs, output and outcome in the health care context as well as the basic approaches towards volume measurement (direct volume measurement, volume measurement via deflation. As the health care sector is made up of very distinct health providers (hospitals, nursing homes, general practitioners, specialists etc.), the ensuing sections of the chapter are organised by type of provider as they give rise to distinct approaches towards measuring output.

4.1.2 Target definition of health care services

4.8. The objective of the present handbook is to provide guidance on the measurement of the volume of health services as provided by the health care industry. It has been pointed out earlier that this is a different focus from investigating the functioning or the effectiveness of the health system as a whole. For example, an improved working of the health system may be noted if there is more preventive activity or if public policy manages to influence people's behaviour so that they adopt more healthy lifestyles. This may very well go hand in hand with a reduced volume of health care services, in particular if such services are of a curative nature.

4.9. In the case of diseases, our central notion in defining health care services is the ***treatment of a disease or medical services to prevent a disease***. Volume measures of output are then disease-based measures. Ideally, in the case of a treatment, the unit of output would capture *complete treatments*, and would take into account quality change in the provision of treatments. This measurement of health care output would then be able to differentiate among price, quantity and quality changes.

4.10. A complete treatment refers to the pathway that an individual takes through heterogeneous institutions in the health industry in order to receive full and final treatment for a disease or condition. This definition of the target measure, otherwise known as a disease-based estimate of health care output, is similar to that used in the Eurostat Handbook (2001), Berndt et al (2001) and Aizcorbe et al (2008), Triplett (2009). We note, however, that the notion of a complete treatment is not always applicable, for instance in the case of chronic diseases. The medical service would then consist of the treatment of one episode of a disease. Similar reasoning applies to the definition of health care services in the area of residential care and nursing homes, more of which below.

4.11. Our target definition of health care services includes medical services to prevent a disease. This has been added to include prevention services that are clearly delivered in the health care setting, such as endoscopy or mammography. However, preventive services that are not of a medical nature such as anti-smoking campaigns are excluded from this definition.

4.12. In addition to curative or preventive disease-related medical services, there are medical services that are not or only partially related to diseases. These are in particular long-term care and nursing services, certain services towards improving states of physical or mental health and possibly some specific interventions such as plastic surgery when carried out for aesthetic reasons. There is no single definition for health care services that would encompass all these aspects in full but it is felt that the target definition put forward is applicable to a large and representative number of cases.

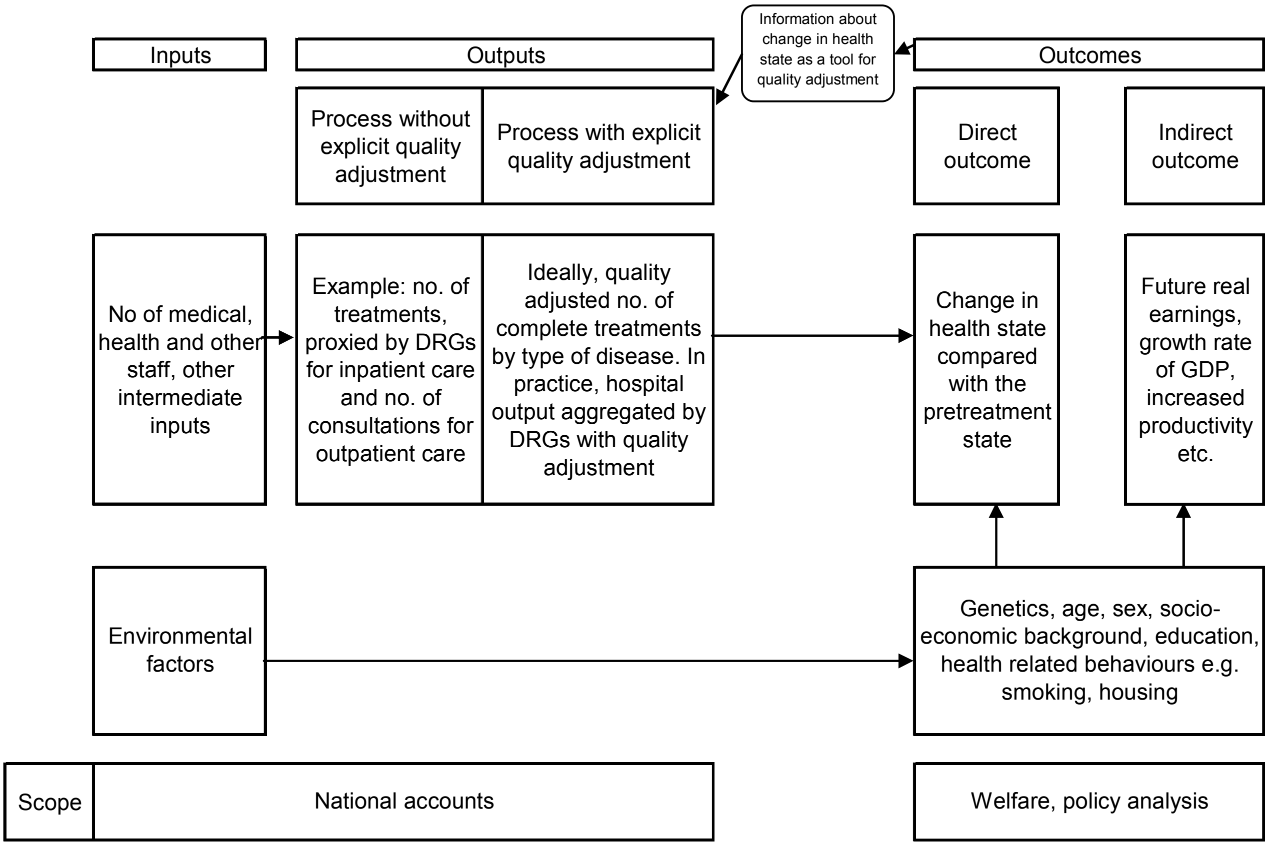
4.13. It will also become evident below that even in the case of disease-related health services, the target definition that does not lend itself directly to implementation in all instances, for practical and for conceptual reasons. But it constitutes a conceptual benchmark that is useful when measurement decisions are made. Before further discussion of the measurement aspects of this definition of health care output, it is useful to put it in relation to the notions of inputs, processes and outcomes (Figure 4.1).

4.1.3 Inputs, processes and outputs

4.14. Inputs comprise the resources used in the production of health services. Inputs include time of medical and non-medical staff, the drugs, the electricity and other intermediate inputs purchased and the services provided by equipment, buildings and other capital used in production.

4.15. Activities – better: processes - in health care arise from the use of resources and are intended to benefit the individual patient. They can be thought of as components of treatments of a particular disease resulting in the provision of health care, which in turn constitutes the output of health care providers. In health services, processes include operative procedures, diagnostic tests, outpatient visits, and medical consultations, individual prevention and counselling. Sometimes, a particular process may coincide with a treatment but this is not generally the case. Nor is the count of processes in itself necessarily a sufficient measure of the output of health care services. This is in particular the case, when processes reflect only part of a treatment, and when processes undergo quality change. However, in a number of cases, the count of processes may be a reasonable or the only practicable approach towards health care service measurement.

Figure 4.1. Inputs, outputs and outcomes: health care sector



4.1.4 Outcome

4.16. Outcome is used in this Handbook to describe a state that consumers value, for example health status. The provision of medical services and its effects on outcome do not necessarily coincide – there may be lags of different length and other factors interacting with medical services. These are in particular:

- Socioeconomic factors such as income, income distribution, employment, education;
- Behavioural factors such as tobacco, diet, exercise, hygiene;
- Environmental factors such as housing, water, pollution;
- Personal factors such as genetics, age and gender. In addition, co-production in health occurs, as the health industry itself does not produce health outcomes but can only support people in realising their potential health. Thus outcomes are also dependent on the individual efforts of patients[17].

4.17. The measurement of the impact of medical intervention on health status or outcome is beyond the scope of this handbook, as the purpose of the national accounts is to measure values and volumes of economic transactions, not the effects or outcomes from transactions. However, in the discussion on quality adjustment below, the point will be made that while outcome as such is not the measurement target of the national accounts, output and outcome are not independent of each other. In particular, statements about quality improvement or deterioration of a medical service cannot be made without some reference to the effects of a medical service on outcome, i.e., to the improvement or deterioration of the state of health of the patient.

4.18. Even if not always relevant for the (core) part of the national accounts, a significant literature is developing in conjunction with measuring the outcome of health care activity. In this framework, health is considered to be a type of human capital which like other capital goods depreciates over time and requires investment. Accounting for the health investment requires inclusion of the value of time that members of households invest in their health (e.g. exercise, sleep, waiting for medicals services) and in the health of others (e.g. home nursing). Other research has also relied on outcomes. There is an increasing number of studies that have used measures of direct outcome to value health care output[18]. For instance, Cutler, McClellan, Newhouse and Remler (1998) derive a price index for heart attack treatment.

4.2 Output: general measurement issues

4.2.1 Value of output

4.19. Throughout this handbook, it is understood that the value of output of institutional units in the health care industry is measured by the observed money value of output in the case of market producers and by the sum of costs in the case of non-market producers. This follows national accounts conventions. Other, more research-oriented approaches exist where the value of production of non-market producers is estimated in other ways than by summing costs and the reader is referred to the relevant literature, for example Nordhaus (2002). But for present purposes, the national accounts convention will be observed.

4.2.2 Volume of output

4.20. The target definition of health care volume output proposed earlier is the number of complete treatments with specified bundles of characteristics so as to capture quality change and new products. This section starts by qualifying this ideal definition in several respects, mainly imposed by data constraints.

Limits to measuring complete treatments

4.21. A first limitation arises with regards to measuring *complete* treatments. In concept, ‘complete’ is understood as a complete treatment pathway across the health care system. As an example of a complete treatment pathway, take a hip replacement operation. The pathway approach would imply aggregating all services or procedures associated with the intervention for the condition whether it is received from primary care services such as a general practitioner, specialists, at hospitals, or at a rehabilitation service. Thus, using the pathway approach would entail collecting data on outputs from a number of health care providers and aggregating them in a meaningful way. This is very challenging.

4.22. There are additional reasons why the principle of complete treatment is difficult to implement in the national accounts:

- In the SNA, total output of an industry is based on summing up outputs of various service providers (establishments), and therefore a complete treatment is hard to capture if service provision cuts across several establishments. Even if it were possible to observe complete

18 See also the volume edited by Cutler and Berndt (2001) for other examples of new medical care price indices.

treatments if there are several service providers involved (e.g. hospitals and outpatient services), there would be no simple way to allocate the overall service to the different participating units and yet this is a requirement for national accounts purposes;

- Most data retrieval systems do not have the capacity to link the treatment of an individual across institutions to enable measurement of the compete treatment. Data on both expenditures (value of inputs) and services received would be required. Thus a health care pathway approach has demanding data requirements as patient records have to be linked across activities and institutions. Even within institutional settings, data may not be appropriately linked[19];

- The beginning and end point of a treatment pathway is observable in the case of acute health conditions but unclear for chronic health problems or for medical conditions that give rise to long-term care and services provided in nursing homes. Many of the diseases associated with ageing and most psychiatric conditions are chronic, long-term conditions, and the patient may be treated for more than one illness or problem within a period. Thus the boundaries of the complete treatment would be unclear;

- An additional complication in the estimation is that pharmaceuticals used in health care are normally included in the total cost of an inpatient service but are a separate activity in outpatient services.

4.23. Given the difficulty with compiling complete treatments, estimates of health care output usually occurs at the institutional level. Thus a narrower view of a treatment is that defined by the type of health service. This measure captures the full treatment only within an institution and generally by function or type of service. Norway, and the UK have for instance adopted this practice. Dawson et al. (2005) compile an output index with 1700 categories of NHS activity including primary care. This aligns with standard practice in national accounting.

A working definition of output

4.24. Thus, rather than reasoning in terms of complete pathways of treatment across the health system, the output measures proposed in what follows are best thought of ***episodes of treatment of particular diseases as provided by a given institutional unit***. Furthermore, this measurement objective will mainly be applicable for curative care whereas other measures will have to be targeted for long-term care and other specialised services where it is difficult to establish when an episode of treatment is complete. For example, inpatients in nursing homes do not generally receive treatment for a specific illness or illnesses where there is an obvious start and end point. For such institutions, a strong case can be made that the output is defined by the processes of the institution of care, not a treatment. The same applies to chronic and progressive health conditions where the patient faces a slow, variable and unpredictable progression of a disease.

4.25. Direct measures of output in these units will have to rely on measures such as the number of particular processes (procedures, consultations etc.) or the number of patients treated in various institutional settings. There are advantages in continuing within this framework until linking of patient treatment across institutions becomes available.

4.26. The present considerations suggest that it is best to treat the measurement of output of medical services by type of health care provider. The most important drawback of this approach is that it is not able to capture substitution effects between providers. For instance, if treatment of a disease moves from a

19. There have been developments in some countries in linking patient utilisation profiles across institutions.

hospital-based, inpatient treatment towards an ambulant or outpatient treatment, this shift and the ensuing consequences for unit costs of output will not be captured. Section 4.3 provides further discussion on this largely unresolved measurement issue.

Implications of separate measurement of inpatient and outpatient treatments

4.27. One specific consequence of the working definition of output above is that it does not permit tracking treatments across institutional units or capturing the effects of shifts between inpatient and outpatient treatments. Outpatient care refers to all medical and health care delivered to individuals when they are not classified as an admitted hospital inpatient. Thus the care may be received in an outpatient facility of a hospital or in a facility not attached to a hospital such as a doctor’s consultation rooms. It is feasible to measure output defined as the number of complete treatments differentiated by type of disease in the acute hospital setting as there are well-developed classification systems. But once treatment moves outside the hospital, the paucity of case-mix classification systems means that measurement often reverts to summing up numbers of processes or numbers of patients. And there have been reports of an ongoing shift between inpatient and outpatient treatment of diseases.

4.28. Furthermore, between inpatient and outpatient care there is an area which includes services known variously as day care, day treatment, or day surgery. These services are provided either at freestanding facilities or in dedicated units within hospitals. They relate to different types of activities/treatments, spanning from rehabilitation therapy to surgery. To measure these elective/planned services, several countries use the same tools that are available for inpatient care. For example, in England, day surgery cases are categorised under the same HRG classification system used for inpatient; moreover, an equivalent tariff/price is posted for the same case type treated as inpatient and day case.

4.29. Note also that pharmaceuticals are an intermediate input in the case of inpatient services whereas they go directly to final consumption in outpatient care. Consequently, in hospitals, changes in the use of pharmaceuticals or their qualitative developments are reflected in intermediate consumption. In the case of outpatient services, the change in the health state of a patient may result separately from both the contribution of the service provider and prescribed pharmaceuticals, but only the former factor is recorded in the health output. The activity of both the pharmaceutical sector and retail pharmacies is recorded outside of the human health services under SNA in both the CPC and ISIC categories. Prescription drugs and other medicines used in either an inpatient or outpatient setting are a part of output of the pharmaceutical products industry.

4.30. Several avenues exist to split current-price values of medical services into a price and a volume component, and partly, the avenue chosen depends on whether services are provided by market or by non-market units. It was explained above that in the SNA, the value of output for market producers is measured by their revenues and the value of output for non-market producers is measured by summing up costs. When it comes to measuring volume (indices), there are two basic options: construction of a direct volume index or deflation of values by a price index.

4.31. In a market-based health system where there is information on market prices or where prices are significant, expenditure on the treatment of a disease can be deflated by a disease-specific price index to arrive at a volume output measure of the disease. For example, Berndt et al (2000) have estimated a price index for heart attacks and this index can be used to deflate disease-specific expenditures. This is similar to what happens in other market sectors in the economy where volume output measurement is accomplished by dividing data on revenues or sales by a price index. Under ideal conditions, the prices for privately provided health goods and services would reflect the marginal costs of production and the marginal utility to consumers.

4.32. In the debate, deflation procedures are therefore often exclusively associated with market producers. This reflects the idea that constructing a price index requires the presence of market prices and the latter are directly associated with market production. While this argument is correct, things are less clear-cut if one allows for a more comprehensive meaning of 'deflation'. In particular, 'deflation' can be understood as applying a true market price index but it can also be understood as applying a unit cost or 'quasi price index' (Schreyer 2008).

4.33. In some countries, hospitals and other providers of medical services are considered market producers because they receive economically significant revenues from reimbursement schemes that, on average, cover their costs. In such cases, a quasi price index consists of average revenues per treatment. One notes, however, that reimbursement schemes are themselves based on cost so that the differentiation between costs and revenues is blurred. Also, the fact that there are revenues does not imply that there is a competitive market where prices necessarily carry signals about consumer preferences.

4.34. Unit costs are the costs per unit of service. As medical services have been defined as the number of (complete) treatments of particular diseases, unit costs are the costs per treatment of a disease such as a heart attack. Note that despite the fact that 'costs' enter the picture, unit costs are defined via outputs (treatments) and not inputs. A unit cost index is therefore a weighted average of unit cost indices of particular diseases, where the cost share of each type of treatment constitutes the weight. Such a unit cost index mimics a price index and can be used for deflation when production is on a non-market basis. Applying a unit cost index to an index of total costs is tantamount to constructing a direct volume index. The unit cost ('cost-per-episode of illness') approach has long been suggested as an option for volume measurement (Scitovski 1967).

4.35. In some instances, it may also be possible to draw on market price information for purposes of deflating values of non-market production. A potential candidate is the medical services part of the Consumer Price Index. However, care has to be exerted to make sure that the CPI is representative for the deflation of the non-market production. In particular,

- The services supplied by the market provider have to be sufficiently similar to those supplied by the non-market provider – this is true for each type of service and for the mix between different services;

- The scope of the CPI has to match the scope of non-market production. This may not be the case when the CPI is designed to reflect prices for out-of-pocket expenditures and when consumers only pay part of the full price for the medical good or service. In this case, the CPI is not an appropriate tool for deflation of non-market production which relies on a concept of measuring production at its full cost.

4.36. Alternatively, direct volume indices can be constructed. A direct volume index is the weighted average of the volume indices of different types of treatments, where the cost share of each type of treatment constitutes the weight. Berndt et al. (p.173) suggest that "real output of medical care could be formed from cost of disease accounts by counting quantities of medical procedures (the number of heart bypass operations, say, or of appendectomies, or of influenza shots), and weighing each procedure by its cost." Although there are some differences between a direct volume index and a volume index derived at by deflation (such as index number formulae, timeliness of data), the basic idea remains the same – volume measures of outputs are sought, as opposed to volume measures of inputs[20].

---

20 An insurance approach has also been considered. For example, Israel considered as a possible indicator of health services the number of insured persons eligible to receive medical services, as determined by agreements between the government and health management organisations. Similar approaches have been considered for

4.37. It is also worth adding that the distinction between market and non-market producers is far from clear-cut. The SNA specifies no unambiguous rule as to what constitutes a non-market producer, and even where such rules exist as in the European Union, their applicability is not straight forward. In OECD countries, there are many shadings of market-based and non-market based health systems and institutions and the allocation of institutional units to the group of market or non-market producers brings with it a certain element of arbitrariness.

4.38. A particular difficulty with health goods and services is that the sector, private or public, is not usually very competitive. Sources of market failure including externalities, moral hazard and imperfect information render the health care market different from markets for other goods and services. Consumers tend to be well-insured, and this places doubt on the extent to which the price they pay reflects their marginal valuation. The nature of health care as a commodity often means it is considered inappropriate to allocate it on the basis of willingness to pay. Hence, decisions about methods for the measurement of volume outputs should be made pragmatically and on the basis of available information rather than on the sometimes tenuous distinction between market and non-market production. What counts for the present purpose is that output measures aim at tracking outputs and not inputs into medical care production.

4.3 Volume output: measurement by provider industry

4.39. The following discussion is organised by provider industry for practical reasons. First, as already emphasised in the previous section, data are not available to measure health volume by disease across health care providers. Second, some health care institutions produce quite different products from others (for example curative treatments of an acute problem versus ongoing treatment for a chronic condition) which makes it natural to discuss output measurement for each industry separately.

4.40. The following discussion will distinguish between hospital activities, residential care activities, medical and dental practice activities and other human health activities, following the ISIC, which provides the building blocks for the production side of the national accounts. For each type of health care activity, the measurement of health volume output requires identifying a set of homogeneous products and a set of weights that can be used to aggregate volume changes in these products (direct volume measurement) or to aggregate unit cost changes in these products (deflation approach – see above).

4.3.1 Hospital activities

4.41. The standard industry classification differentiates between acute hospitals, mental health and substance abuse hospitals and speciality hospitals. For purposes of output measurement, it will be useful to keep this distinction because there are good reasons to believe that services provided by these units are different.

Acute hospitals

4.42. Typically, acute hospitals provide inpatient as well as ambulant treatments. Within a hospital, different outputs can be captured by identification of treatments. Although this may not always be possible in practice it is desirable from a conceptual viewpoint not to separate inpatient and ambulant treatments of the same disease so as to be able to capture effects of substitution between inpatient and ambulant treatments.

private health care. Obviously, this is only an indirect approach towards measuring health services provided and in many ways closer to an input-based than to an output-based measure. Also, no account can be made of productivity change in health provision.

4.43. To measure treatments, there are two common patient classification systems both of which attempt to deal with the heterogeneity of hospital output while making comparisons between hospitals possible. The first system is the international classification of diseases (ICD) which was originally developed as a basis for mortality statistics. Thus it refers to diagnoses. It is used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and hospital records. The ICD underlies the development of DRG categories. Responsibility for updating ICD lies with the WHO and the ICD-10 was endorsed in 1990.

Use of DRGs

4.44. The most widely available categorisation of hospital inpatient services is provided by diagnosis related groups (DRGs). They were developed with the explicit objective of creating relatively cost-homogeneous groups in order to compare hospital performance. Instead of providing a cost for each component of a hospitalisation, DRGs give a composite bundle of hospital services a single predetermined cost or reimbursement rate. This amount includes all activities from which the patient benefits in the process of the treatment, including nursing care, drugs, imaging and the hotel amenities of care. The main characteristics of the DRG system are described in Annex B.

4.45. DRG systems are attractive for the measurement of volume output because they provide information on (unit) costs per type of treatment and on the number of treatments carried out. In other words, there are the basic ingredients for a (quasi) price or a volume index.

4.46. A typical DRG system comprises a large number (500-1000) of categories. By construction, each category stands for a relatively homogenous service and thus, in principle, construction of a unit cost or of a volume index from the most detailed level of categories is desirable. This is, however, not always possible. A main reason is that DRG systems are updated on an ongoing basis with some categories being aggregated and others disaggregated, making comparisons between periods difficult. To deal with this problem, DRG categories are sometimes grouped into broader diagnosis groups as explained in the example from Germany (see Box).

4.47. Given cost weights and the number of treatments, either a unit cost (quasi-price) index or a direct volume index can be constructed. The choice between these alternatives is often governed by data availability. In many countries, information about the evolution of average costs per treatment may be more quickly available than information about the evolution of the number of treatments. When indices are constructed such that the weight reference period precedes the latest period for which (quasi) price changes are to be measured, a unit cost index is easier to construct than a direct volume index.

Hospital discharge numbers

4.48. Hospital discharge registers constitute an alternative data source for the construction of a volume index of hospital output. Discharges correspond to diagnoses for particular diseases and can be grouped in a meaningful way so as to represent similar types of episodes of treatment. For example, Statistics Netherlands groups discharge data into about 1 000 diagnostic groups, on the basis of the ICD-9 classification. Some groups are further stratified by age of patient or severity of disease so as to obtain relatively homogenous clusters of diseases.

Box 7. A hospital price index on the basis of DRGs in Germany

In Germany, for example, the DRG system covers over 1000 categories. For the construction of a unit cost index, the German Federal Statistical Office identifies 27 major diagnosis groups which then serve as the components of the resulting price index. Each DRG category carries with it a cost weight that depends on the medical complexity of the treatment associated with a particular DRG. In Germany, these cost weights range between 0.106 (contractions, 1 day in hospital) and 64.899 (transplantation of liver, lungs, heart or stem cells, artificial respiration of more than 999 hours). Cost weights are valid for one calendar year. Along with cost weights comes a base rate, the reference price that a hospital can charge for an average DRG. Thus, for a DRG with a cost weight of 1.0, the hospital charges exactly the base rate which in 2008 is around € 2 800. Note that 'weight' here is not used in the sense of a set of shares that sum to unity but rather in the sense of an adjustment coefficient. In what follows, the term 'coefficient' shall therefore be used.

DRGs are then aggregated into 27 major diagnosis groups. This avoids a problem with changing grouping of treatments between DRGs that arises at lower level of disaggregation. The next step consists of computing average adjustment coefficients for each major diagnosis group. Denote with c_i^t ($i=1,\ldots N_j^t$) the coefficients for each of the N_j^t different DRGs in period t and denote with x_i^t the number of corresponding cases. The average coefficient for each of the j = 1,…27 major diagnostic groups is then given by

$$c_j^t = \frac{\sum_{i=1}^{N_j^t} x_i^t c_i^t}{\sum_{i=1}^{N_j^t} x_i^t} = \sum_{i=1}^{N_j^t} \frac{x_i^t}{\sum_{i=1}^{N_j^t} x_i^t} c_i^t$$

c_j^t is thus a weighted average of the coefficients of individual DRGs in the major diagnostic group where the share in the total number of treatments, $\frac{x_i^t}{\sum_{i=1}^{N_j^t} x_i^t}$, constitutes the weights. Average revenues per major diagnostic group in period t are measured by multiplying average adjustment coefficients c_j^t by the base rate BR^t applicable for period t.

$$AR_j^t = c_j^t BR^t.$$

In the final step in the computation of the price index between two periods t and t-1, the changes in unit costs per major diagnostic group are weighted by the revenue share of each group where $C_j^t = AR_j^t x_j^t$ is the total revenue (reimbursement) that the hospital receives for major diagnostic group j.

$$\text{Price index}^{t,t-1} = \sum_{j=1}^{27} \left[\frac{C_j^{t-1}}{\sum_j C_j^{t-1}}\right]\left[\frac{AR_j^t}{AR_j^{t-1}}\right].$$

It is interesting to note that the reimbursement mechanism for hospitals implies that the total sum of revenues or reimbursements is ex-post not necessarily equal to costs. Individual establishment but also the hospital sector as a whole may generate a certain profit or a loss for a given period because base rates and cost weights are established ex-ante. On the whole, such differences would, however, be expected to remain contained. This justifies a set-up where in Germany's national accounts, the sum of costs determines the value of production for the hospital sector whereas a (quasi) price index based on revenues determines the movement of volumes.

Source: Based on Statistisches Bundesamt (2008).

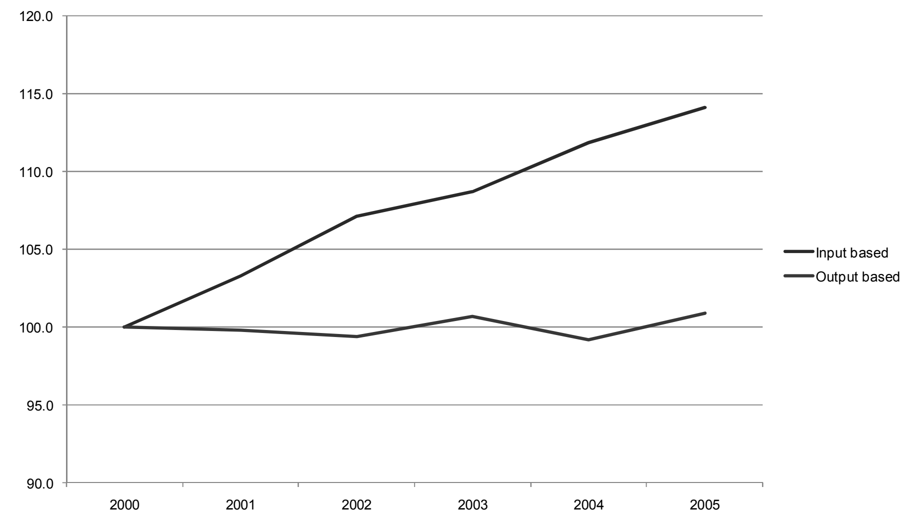
4.49. Unlike DRG-based information, hospital discharge records are not normally accompanied by cost information. Absent such data, proxies have to be formed to weight changes in discharge numbers by diagnostic group. In the Netherlands, the share in total hospitalisation days for each diagnostic group is the proxy employed. In other words, the implicit assumption is made that the unit costs per treatment are proportional to the average number of hospital days per treatment, with a constant proportion across treatments.

4.50. One advantage of using hospital discharge rate information is that, for a given disease, no distinction is made between inpatient and day treatments. This helps dealing with substitution processes, such as shifts from inpatient to ambulant treatments for certain diagnoses. For a given weighting structure, such a shift would have no volume effects – as should be the case – and any movements in value will be captured as price changes. Had ambulant treatments been kept separate, and thereby treated as separate products, a shift from inpatient to ambulant treatment would have resulted in a measured reduction in volume output – a counter-intuitive result.

Box 8. A hospital price index on the basis of DRGs in Denmark

In Denmark, the DRG system contains information on about 800 categories of treatment, on associated fees and on the number of treatments. Unlike Germany, Denmark constructs its (quasi) price index directly from individual DRG categories, after excluding categories that are not comparable across adjacent periods. The index is computed for acute hospitals, i.e., excluding psychiatric hospitals. The figure below compares the resulting (quasi) price index with the input price index for hospitals traditionally used for deflation in the national accounts. The results show that, over the period 2000-05, the input price index rises quicker than the output-based index, implying faster volume growth when the output-based index is applied to hospital costs.

Input price index and output-based price index for acute hospital services in Denmark



Source: Deveci, Heurlén and Sorensen (2008).

Mental health and substance abuse hospitals

4.51. In principle, mental health and substance abuse hospital services lend themselves also to a DRG-based measurement of output. DRG-based indices are, for example, used in Austria. However, in many other countries, psychiatric hospital treatment is not or only partially covered by the DRG system and so other methods must be used for output-based indicator development. This less favourable data situation is somewhat attenuated by the fact that psychiatric hospitals tend to account for a moderate share of total hospital costs, so any measurement error enters with an equally moderate weight.

4.52. In the absence of true output-based measures of production, substitutes include discharge rates weighted by days of hospitalisation (see above), or even simpler indicators such as number of patients or un-weighted numbers of discharges. Obviously, all these indicators suffer from the fact that no account is taken of the severity of diseases, and the type of treatment required. Consequently, they can only qualify as proxies to output-based measures of production.

4.3.2 Residential care activities

4.53. It has been mentioned earlier that for residential care activities, the notion of a complete treatment is not very meaningful. Elderly and other long-term stay patients tend to have complex clinical presentations characterised with disability, dependency and multiple pathologies. The start and end point of the condition is not clear and the fact that the condition is chronic means that incentives for reducing length of stay as under a DRG system would be inappropriate. Annex 4.C provides a description of the RUG system. A classification for nursing home patients, Resource Utilisation Groups (RUGs) has been developed, as DRGs are of little value for chronic care patients and those without straightforward clinical conditions.

4.54. Where countries do not have a RUG-type system, the number of occupant days differentiated by the level of care can be used. Many countries have their own care classifications based on an assessment of a patient's care needs and thus the intensity (and cost) of the care received. If information on costs per level of intensity of care is available, it can be combined with data on the number of occupant days to derive an approximate direct volume index of output or to derive an approximate unit cost index.

4.55. This is, for instance, the case in Denmark. Data on unit costs for different types of residential places (nursing homes, sheltered housing, day centres and social centres) exist, if only for the city of Copenhagen. Combined with the evolution of the number of care places, grouped in the same way, this is used to construct a unit cost index for residential and care places.

4.56. When residential care facilities are considered to be market producers, a specialised price index, for example a CPI component, can be used for deflation. When non-market producers provide residential care services comparable to market producers, it is also possible to apply the specialised market price index to deflate costs of non-market producers. However, care must be taken to make sure that CPI definitions and its scope are meaningful for deflating non-market service output.

4.3.3 Medical and dental practice activities

4.57. Medical and dental practice activities are considered market production in many countries. Persons receiving dental services in particularly are more likely to be charged an economically significant price. Both medical and dental practice activities provide both general and specialist services. In the case of both types of services an appropriate deflator for market output would be a price index such as a CPI component that accounts for different types of services received and captures – to the extent possible - quality changes.

4.58. As a rule, the definition and measurement of outpatient treatments remain rudimentary (Castelli, 2007). A limited number of countries have developed and use outpatient classification systems. In the same way as for inpatient activity, all outpatient activity related to one treatment episode would ideally be combined into one measure of output. The episode would include all consultations, pathology tests, imaging and prescriptions. Development of outpatient DRGs requires a capacity to track patients across outpatient services for the same treatment. To do so, it would be necessary to be able to identify the start and end point of the complete treatment and to have an appropriately supportive legal and information technology framework. There have been limited developments in outpatient DRGs, some of which are referred to in Annex 3.

4.59. Until an international or more widespread national classification systems for outpatient care are developed and implemented, basic quantity measures such as number of doctor visits etc. will have to be used in the construction of volume measures, in particular for non-market providers. The EU Handbook on Prices and Volumes in the National Accounts suggests that outputs should be classified into medically meaningful groups that are as a homogeneous as possible. The stratification may take into account the medical content of the output as well as a time dimension. For example, a visit to a general practitioner could be a measure of output. Other quantity indicators would be patient transfers by ambulance, number of pathology tests by broad category of type of test, number of prescriptions filled by type. Generally, these data are collected as part of the process of reimbursement either publicly or privately.

4.60. An example for a simple quasi-price index for non-dental practice activities comes from Austria where social security information on reimbursement, along with the number of cases, provides the components for a deflation method. Doctors are grouped into 18 groups of specialists. Then, average turnover per case is computed on the basis of total revenues of physicians paid by social security funds and the number of cases treated. Changes in the average turnover per case and group are weighted together with the corresponding revenue shares of each medical group. This forms the price index.

4.61. The EU Handbook suggests that general practitioners consultations are measured by number of visits but specialist consultations are measured by the first visits only. The reason given for this differentiation is that specialists' visits are more likely to be follow-up visits, i.e. ongoing treatment for the same medical condition. This distinction seems arbitrary, as many GP visits are also follow-up visits. In addition, while this recommendation may be applicable to some specialties, it may not be applicable to all. Specialties to which the notion of ongoing treatment would not usually apply include many diagnostic specialties such as pathology, radiology, nuclear medicine etc.

4.3.4 Other human health activities

4.62. This category refers to a range of diverse activities such as activities of nurses, midwives, physiotherapists or other paramedical practitioners in the field of optometry, hydrotherapy, medical massage, occupational therapy, speech therapy, chiropody, homeopathy, chiropractics, acupuncture etc. Many of these services are provided by market producers.

4.63. In the case where some of the activities under Other Human Health do not have significant prices, e.g. blood and organ donation, it will be necessary to aggregate the output by using relatively basic methods such as the number of consultations, visits or tests performed.

4.3.5 Overview of measures

4.64. To complete the discussion above, Table 4.1 provides a summary of suggested output measures by ISIC categories. Only producers under the ISIC rev 3.1 category Division 85 "Health and Social Activities" or ISIC rev 4 are included. That is, retail pharmacies, health insurance and other production activities outside of that classification are not considered. The proposed methods are in principle output-based, although the degree to which they constitute full-fledged measures of output varies between them and sometimes depends on the level of stratification in implementation. No explicit mention is made of quality adjustment but it should be well understood that quality adjustment - if feasible and applied with care - would lead to further improvement of the methods mentioned.

Table 4.1. Overview of indicators for volume output of health service providers

	ISIC rev 3.1 & 4	Output-based methods
Hospital activities		
Acute Hospitals	8511 & 8610	(Quasi) Price index based on DRGs (cost or revenue-weighted) Direct volume index based on DRGs (cost or revenue-weighted) Direct volume index based on ICD categories (e.g., number of discharges by category with quantity-weights such as shares in hospital days)
Mental health and substance abuse hospitals	8511 & 810	(Quasi) Price index based on DRG-like categories (cost or revenue-weighted) Direct volume index based on DRG-like categories (cost or revenue-weighted) Direct volume index based on ICD categories (e.g., discharge numbers with quantity-weights such as shares in day care days) Number of discharges* Number of days of care*
Speciality (other than HP.1.2) hospital	8511 & 8610	
Residential care activities		
Nursing care facilities *Note: RUGS are only used for nursing care*	8519/8531 & 8710	(Quasi) Price or unit cost index based on Resource Utilisation Groups (RUGs) or equivalent (cost-weighted) Direct volume index based on RUGs or equivalents (cost-weighted) Direct volume index based on number of days of care by level of care (cost weighted) Direct volume index based on number of cases by level of care (cost weighted) Number of days of care* Number of cases/discharges*
Residential mental retardation, mental health and substance abuse facilities	8519/8531 & 8720	
Community care facilities for the elderly	8519/8531 & 8730	
All other residential care facilities	8519/8531 & 8790	
Medical and dental practice activities		
Doctor services *Note: services are defined as consultation/visit/treatment depending on the typology of the country*	8512 & 8620	(Quasi) Price index based on number and type of service (cost or revenue-weighted) Direct volume index based on number and type of service (cost or revenue-weighted) Relevant component of Consumer Price Index if applicable** (Quasi) Price index based on average costs/revenues per service (cost or revenue-weighted) Direct volume index based on number of services (cost or revenue-weighted) Number of services*
Dental services *Note: 'number of services' refers to units such as consultations, visits or treatments, depending on the typology of the country*	8512 &8620	Relevant component of Consumer Price Index if applicable** Direct volume index based on number of services (cost or revenue-weighted) Number of services*

	ISIC rev 3.1 & 4	Output-based methods
Other human health activities *Note: the list of services below is not exhaustive as other human health activities covers very heterogeneous activities*		
Other health practitioner consultations	8519 & 8690	Direct volume index based on number of consultation by type of consultation (cost or revenue-weighted) (Quasi) Price index based on average cost or revenue per consultation (cost or revenue-weighted) Relevant component of Consumer Price Index if applicable** Number of consultations* Number of tests performed* Number of cases treated*
Other outpatient visits	8519 & 8690	
Family Planning centres	8519 & 8690	
Outpatient mental health and substance abuse centres	8519 & 8690	
Free-standing ambulatory surgery centres	8519 & 8690	
Dialysis care centres	8519 & 8690	
Other outpatient multispecialty and cooperative service centres	8519/8531 & 8690	
All other outpatient care centres	8519/8531 & 8690	
Medical and diagnostic laboratories	8519 & 8690	
Home health care services	8519/8531 & 8690	
All other ambulatory health care services	8519 & 8690	
Ambulance service	8519 & 8690	
Blood and organ banks	8519 & 8690	
All other ambulatory health care services	8519 & 8690	

*: proxy index.

**: note qualifications on the use of the CPI above.

4.4 Quality adjustments

4.4.1 Capturing quality change via stratification

4.65. Price and volume measures of output should reflect quality changes in the health services provided. In other words, only prices, unit values or quantities of the same quality, i.e., with the same characteristics should be compared over time. It has been noted earlier that a first and important step towards capturing quality change is the correct stratification, i.e., a comparison of products with the same or at least similar characteristics. In this way, stratification keeps quality constant if the products included in a particular stratum are relatively homogenous.

4.66. An example in health services would be the matching of hospital services. Public and private services may provide the same treatments and yield the same health outcomes but if the amenities provided are valued differently by consumers, the services should not be matched. In addition in some countries, consumers use private hospitals because there may be a waiting time for a public hospital procedure but no waiting for the same procedure in a private hospital. Thus under these circumstances, private hospital services may not be considered as a substitute for public hospital services and services provided in the two types of units should be considered different products.

4.4.2 Explicit quality adjustment

4.67. Matching of services has its limits when comparable products do not exist in comparison periods or when new services only gradually diffuse in practice (see Box 2, Chapter 1). This is the case even when using relatively sophisticated output measures such as DRGs and requires, in principle, explicit quality adjustment.

4.68. Quality of health services is multifaceted. It can relate to subjective perceptions of patients, it can relate to the extent to which health services increase the likelihood of desired health outcomes and it can relate to the extent to which the right choices are made about procedures and treatments, i.e. to the extent to which medical care is consistent with current professional knowledge (Lohr, 1990).

4.69. Process quality is generally judged according to whether the right choices are made in treating the patient by assessing adherence to professional standards. The professional standards of care or guidelines are developed from clinical trials or clinical evidence and the quality of the process is assessed by considering the compliance of medical practice with the evidence base. Examples of process quality indicators are rates of influenza vaccination for adults over 65, rates of retinal exams in diabetics and colon cancer screening rates. New treatments and improvements in practice are only incorporated into the guidelines for treatment when the evidence base that the new or improved treatment leads to improved outcomes is sufficiently compelling and well-established. Evidence based medicine has the potential to reduce variations in care, decrease resource utilisation as well as improve health care quality.

4.70. In an ideal situation, explicitly quality-adjusting volume output for process quality would require an adjustment factor reflecting the compliance rate with established procedures by country and disease group. Changes in the proportion, either positive or negative, would indicate where medical practice and procedures have changed to reflect the introduction of new treatments and improvements in the existing practices. There is a limited literature reporting such proportions. For example, Schuster et al. (2005) find that by averaging the findings from US preventive care studies that 50% of people received recommended care, and for acute care studies found that 70% of patients received recommended care and 30% of patients received contraindicated acute care. It is important to note that achievement of full compliance to guidelines, that is a rate of 100%, is not necessarily a public health goal. In industrial production processes, uniformity ensures highest quality of outcomes. The same does not hold in health care as individual differences and preferences need to be accommodated.

4.71. There is a wealth of information on clinical practice guidelines by country but insufficient summary information at this stage on the rate of compliance to best practice. In the USA, the National Guideline Clearinghouse maintains a catalogue of high quality guidelines published by various organisations (mostly professional physician organisations). In the United Kingdom, clinical practice guidelines are published primarily by the National Institute for Health and Clinical Excellence (NICE). In The Netherlands, the Dutch Institute for Healthcare Improvement (CBO) and the Dutch College of General Practitioners (NHG) have guideline development programs that use an evidence-based approach. In Germany, the Agency for Quality in Medicine coordinates a national program for disease management guidelines. All these organisations are members of the Guidelines International Network, an international not-for-profit association of organisations and individuals involved in clinical practice guidelines.

4.72. The second aspect of quality relates to the impact of health services on health outcomes. Health services researchers recommend using both process and outcome indicators for two reasons. First, there is a difference between evidence in research (efficacy) and the outcomes in real life (effectiveness). Second, there is frequently a considerable time period between the process and its impact on the outcome. For example, studies have examined family doctors' compliance with guidelines for hypertension treatment. The outcome associated with this practice, is reductions in AMIs (heart attacks) and stroke (as an intermediate outcome) and mortality related to cardiovascular diseases (as an ultimate outcome). Reductions in the incidence of these diseases occur over a very long time period and hypertension treatment is only one factor involved in the ultimate outcome. Aside from other things patients with hypertension have to comply themselves with a lifestyle involving healthy nutrition and adequate exercise.

4.73. This discussion emphasises the use of quality adjustment using process and outcome indicators. It is also noted that there is an 'industry of quality measures' but at this stage none appear appropriate as a general international recommendation for quality adjustment of health volume output. The main reasons for that lack of applicability is that many of the available process and outcome indicators are country or even institution specific. This should however not discourage countries to continue working towards explicit quality adjustment methods and possibly apply them on a national basis.

4.74. Since quality is multidimensional, it would be ideal to subsume several characteristics of quality into a single indicator that reflects the contribution of the product to outcome. Alternative means have been suggested to derive a single indicator. The first is to choose the most important dimension only, e.g. 30 day survival rate[21]. The second is to use indicators from more than one quality dimension and weigh them as equally important. Third, indicators from a number of quality dimensions can be used but expert opinion should be sought on the appropriate weights (ONS, 2008). Finally, measures such as quality-adjusted life years (QALYs - see Annex) can be used that reduce modifications in health outcomes due to medical care to one dimension, the quality-adjusted gain in time. However, many of these empirical methods are still in a research stage.

4.75. There are a number of desirable characteristics of indicators which could be used for quality adjustment for volume output for determining the marginal contribution of the health industry to outcome. These are outlined below:

- The quality measure should be aligned with the processes sought by consumers, which would generally be a complete treatment by disease;
- The adjustment in output should reflect the marginal contribution of the health industry to an outcome. It should not be affected by any other factors that influence health outcomes such as genetic background, income or lifestyle;
- Consumers are ultimately concerned to achieve an improvement in their health outcome. Waiting times and comfort are secondary to improvements in health status. This points to a conclusion that different dimensions of quality should not be given the same weight[22];
- In many health treatments or processes, there is a time lag before the improvements in health status. Quality adjustment needs to address in a realistic manner the impact of lifetime effects of health expenditures;
- The quality measure should reflect as closely as possible the normal, average or expected effect of the activity on the state of health. Individual capacities to benefit from treatment, or what is known as co-production, should not be counted in the measure of quality adjusted health volume output;
- International comparison is important, and the indicators and methods of output adjustment should be standardised across countries to facilitate comparisons (Smith and Street, 2007).

21 Christian (2007) proposes the use of survival rates at the time of hospital discharge for quality adjustment of hospital volume output. He draws heavily on the work of Dawson et al (2005).

22 There may of course be a connection between some dimensions. For example, a shorter waiting time can lead to improved health outcomes.

4.76. To sum up, methods for quality adjustment of output are still under development[23]. Some headway towards capturing quality can be made by using detailed product specifications and follow the associated costs and treatments over time. The importance of explicit quality adjustment is undeniable but until there is a consensus on techniques for adjustment, it will be difficult to put forward a recommendation for an explicit quality adjustment of health volume output in the national accounts. A similar conclusion has been reached by the United States Bureau of Economic Analysis (BEA) who note that the “BEA will not attempt to account for potential changes in the quality of treatments, a problem where no clear consensus exists on a solution” (Aizcorbe et al, 2008, p. 25).

[23] See for example, Castelli et al (2007)

ANNEX 4.A: CLASSIFICATIONS

4.77. Table4.2 presents a comparison of the content of ISIC, CPC and ICHA-HP (International Classification for Health Accounts health care providers) categories or classes. The content of the ISIC and CPC categories lines up well. The scope and detail of health care used in national accounting diverges with that used in the ICHA-HP and thus under the System of Health Accounts (SHA).

Table 4.2. International categories of health care according to ISIC, CPC and ICHA-HP

ISIC rev 4	ISIC rev 3.1	CPC rev 1.1	ICHA-HP
Section Q: Human health and social work activities	Section N division 85: Health and social work	931: Human health services	Health providers
8610: Hospital activities	8511: Hospital activities	93110: Hospital services	HP.1: Hospitals HP.1.1: General hospitals HP.1.2: Mental health and substance abuse hospitals HP.1.3: Specialty hospitals (other than those in 1.2)
8620: Medical and dental practice activities	8512: Medical and dental practice activities	93121: General medical services 93122: Specialised medical services 93123: Dental services	HP.3: Providers of ambulatory care HP.3.1: Offices of physicians HP. 3.2: Offices of dentists
8690: Other human health activities	8519: Other human health activities	93191: Deliveries and related services, nursing services, physiotherapeutic and paramedical services 93192: Ambulance services 93193: Residential health facilities services other than hospital services 93199: Other human health services n.e.c.	HP. 3.2: Offices of other health practitioners HP. 3.4: Out-patient care centres (also under 8531) HP.3.5: Medical and diagnostic laboratories HP.3.6: Home health care services (also under 8531) HP. 3.9.1: Ambulance services HP. 3.9.2: Blood and organ banks HP. 3.9.9 All other ambulatory health care services
87: Residential care 8710: Nursing care facilities 8720: Residential care activities for mental retardation, mental health and substance abuse 8730: residential care for the elderly 8790: all other residential care	8519: Other human health activities	93191: Deliveries and related services, nursing services, physiotherapeutic and paramedical services 93193: Residential health facilities services other than hospital services 93199: Other human health services n.e.c.	HP.2: Nursing and residential care facilities (also under ISIC 3.1 8531*) HP.2.1: Nursing care facilities 8519/8531 HP.2.2: Residential mental retardation, mental health and substance abuse facilities HP.2.3: Community care facilities for the elderly 8519/8531 HP.2.9: All other residential care facilities

Table 4.2 International categories of health care according to ISIC, CPC and ICHA-HP (continued)

4772: Retail sale of pharmaceutical and medical goods, cosmetic and toilet articles in specialized stores 4773: Other retail sale of new goods in specialized stores	5231: Retail sale of pharmaceutical and medical goods, cosmetic and toilet articles 5239: Other retail sale in specialized stores	62273: Specialized store retail trade services, of pharmaceutical and medical goods 62274: Specialized store retail trade services, of surgical and orthopaedic instruments and devices	HP.4: Retail sale and other providers of medical goods HP.4.1: Dispensing chemists 5231 HP.4.2: Retail sale and other suppliers of optical glasses and other vision products HP.4.3: Retail sale and other suppliers of hearing aids HP.4.4: Retail sale and other suppliers of medical appliances (other than optical goods and hearing aids) (other than optical goods and hearing aids) HP.4.9: All other miscellaneous sale and other suppliers of pharmaceuticals and medical goods
			HP.5: Provision and administration of public health programmes
412: Regulation of the activities of providing health care, education, cultural services and other social services, excluding social security 8430: Compulsory social security activities 6512: Non-life insurance	7512: Regulation of the activities of agencies that provide health care, education, cultural services and other social services, excluding social security 7530: Compulsory social security activities 6603: Non-life insurance	9112: Administrative services of agencies that provide educational, health care, cultural and other social services, excluding social security services 9131: Administrative services of sickness, maternity or temporary disablement benefit schemes 71320: Accident and health insurance services	HP.6: Health administration and insurance HP.6.1: Government administration of health 7512 HP.6.2: Social security funds 7530 HP.6.3: Other social insurance HP.6.4: Other (private) insurance 6603 HP.6.9: All other health administration
			HP.7: All other industries (rest of the economy) HP.7.1: Establishments as providers of occupational health care services HP.7.2: Private households as providers of home care HP.7.9: All other industries as secondary producers of health care

* ISIC 8531 social work activities with accommodation and also CPC 93311 welfare services delivered through residential institutions for elderly persons and person with disabilities.

4.78. The ICHA-HP classification is a refined and modified version of the health-relevant parts of ISIC (rev. 3.1) which was developed to serve the purposes of national health accounting generally. ICHA-HP is broader than ISIC as it focuses on aggregating all health expenditure which may have an impact on health status. There are a number of important differences between the ICHA-HP, and ISIC and CPC classifications. First, ICHA-HP includes more detailed descriptions and a substantially longer list of health care providers than is provided by ISIC (version 3.1 under section N division 85 and version 4.0 under section Q) or CPC under class 931 and used in SNA. The ISIC classifications represent the core institutions of the health care sector. Hospital, doctor and dental consultations and associated activities such as ambulance services, nursing services, physiotherapeutic and paramedical services all perform health related activities as their core activity. ICHA-HP also includes retail sales of medical goods and medicines in pharmacies whereas under ISIC (3.1), this activity is included in 5231 (under retail sale of pharmaceutical and medical goods, cosmetics and toiletries). ICHA-HP also includes a category for private health insurance which is classified in ISIC (3.1) in 6603 (non-life insurance). Additionally, households are recognised in ICHA-HP for their role in the provision of health services. In the SNA, the classification of non-market products covers government and NPISH production, only. In the ICHA-HP, private householders as providers of home care are included as producers (HP.7.2). In this case, social transfers to households caring for patients at home are included as expenditure on health care.

4.79. Another issue is that the ICHA-HP classification, hospitals HP1, includes both inpatient and outpatient services. In the ISIC classification, hospitals (ISIC 8511) provide chiefly inpatient services and, thus, the definition does not exclude the possibility that some services may be for outpatients. However, in both the ISIC and CPC classifications, services of *hospital outpatient clinics* are included under CPC 93121, 93122 or 93123 (medical services) or ISIC 8512 (8620) medical and dental practice activities.

4.80. A problem with the ISIC revision 3.1 has been addressed in the revision 4.0. In revision 3.1, many of the providers under the ICHA-HP classification cross the boundary of health care as defined by ISIC and CPC. For example, community care facilities of the aged are classified as HP. 2.3 under ICHA-HP but as 8519 (other human health activities) and 8531 under ISIC (3.1). The latter category of *Social work activities with accommodation* covers provision of care for all social groups (the aged, handicapped, homeless, and children) together. Thus, there was no distinction between provision of care for the aged and handicapped, and provision of care for other social groups.

4.81. A new item has been added to the ISIC revision 4.0 of 87 *Residential care activities*, under Section Q *Human health and social work activities.* This division includes the provision of residential care combined with either nursing, supervisory or other types of care as required by the residents. Facilities are a significant part of the production process and the care provided is a mix of health and social services with the health services being largely some level of nursing services. This revision in ISIC (rev 4.0) provides better possibilities to create a correspondence between ISIC and ICHA-HP (ICHA Classification of Health Care Providers) at the division level, and to make it possible to recognise long-term care in an internationally comparable way.

Table 4.3. Cross-classification of ICHA-HC and COICOP, COFOG and COPNI

ICHA	Function of health care	COICOP Households	COICOP NPISHs	COICOP Government	COFOG	COPNI
	ICHA-HC function is mainly part of SNA93 code:					
HC.1	Services of curative care					
HC.1.1	In-patient curative care	06.3	13.2.7	14.2.7	07.3	02.3
HC.1.2	Day cases of curative care	06.3	13.2.7	14.2.7	07.3	02.3
HC.1.3	Out-patient curative care	06.2	---	---	07.2	02.2
HC.1.3.1	Basic medical and diagnostic services	06.2.1	13.2.4	14.2.4	07.2.1	---
HC.1.3.2	Out-patient dental care	06.2.2	13.2.5	14.2.5	07.2.3	02.2.2
HC.1.3.3	All other specialised health care services	06.2.1	13.2.6	14.2.4	07.2.3	---
HC.1.3.9	All other out-patient curative care	06.2.3	13.2.6, (13.2.4)	14.2.6, (14.2.4)	07.2.4, (07.2.1)	02.2.2
HC.1.4	Services of curative home care	06.1.2, (06.1.3)	13.2.4, (13.2.7)	14.2.6	07.2.4, (07.3)	02.2
HC.2	Services of rehabilitative care	---	---	---	---	---
HC.2.1	In-patient rehabilitative care	06.3	13.2.7	14.2.7	07.3	02.2.3
HC.2.2	Day cases of rehabilitative care	06.3	13.2.7	14.2.7	07.3	02.2.3
HC.2.3	Out-patient rehabilitative care	06.2.3, (06.2.1)	13.2.6, (13.2.4)	14.2.6, (14.2.4)	07.2.4, 07.2.1	02.2.3
HC.2.4	Services of rehabilitative home care	06.2.3	13.2.6	14.2.6	07.2.4, (07.3)	02.2.3, (02.3)
HC.3	Services of long-term nursing care					
HC.3.1	In-patient long-term nursing care	06.3	13.2.7	14.2.7	07.3	02.3
HC.3.2	Day cases of long-term nursing care	06.3	13.2.7	14.2.7	07.3	02.2.3
HC.3.3	Long-term nursing care: home care	06.3, (06.2.3)	13.2.7, (13.2.6)	14.2.7, (14.2.6)	07.3, (07.2.4)	02.2.3, (02.3)
HC.4	Ancillary services to health care					
HC.4.1	Clinical laboratory	06.2.3, (06.2.1)	13.2.6	14.2.6	07.2.4	02.2.3
HC.4.2	Diagnostic imaging	06.2.3, (06.2.1)	13.2.6	14.2.6	07.2.4	02.2.3
HC.4.3	Patient transport and emergency rescue	06.2.3, (06.3)	13.2.6, (13.2.7)	14.2.6, (13.2.7)	07.2.4	02.2.3, (02.3)
HC.4.9	All other miscellaneous ancillary services	06.2.3	13.2.6	14.2.6	07.2.4	02.2.3

4.82. Table 4.3 shows the cross classification between ICHA-HC (International Classification for Health Accounts health care functions) and SNA 93 classifications. The Health Care classification accounts for the final consumption of health care and the table shows how final consumption is distributed across the three classifications of expenditure according to purpose. COICOP24 is used to classify individual consumption expenditure of households, NPISHs and general government. COPNI and COFOG are used to classify a range of transactions, including outlay on financial consumption expenditure, intermediate consumption, gross capital formation and capital and current transfers by NPISHs and general government respectively.

ANNEX 4.B: OVERVIEW OF COUNTRY PRACTICE

4.83. This annex aims at summarising country practice in the area of measuring health services. It starts out with an overview table that indicates measurement approaches used in OECD countries. A distinction is made between measures presented used and planned measures. The second part of this annex provides descriptions of some country practices for purposes of illustration and further reference.

Austria[25]

General description

4.84. Statistics Austria uses a deflation method for estimation of health volume output. The deflation method is applied to output under OCPA 85 (corresponding to ISIC 3.1 code 85) of hospitals services (code 85.11), medical practice services (85.12), dental practice (85.13) and other human health services (85.14). Almost all groups are classified as market output (non-market output accounts for about 4 to 5 %). The deflation method entails deflating market output by specialised (quasi) price indexes.

Hospital services

4.85. To deflate the subgroup hospital services, the services are grouped into homogeneous groups based on the characteristics of the kind of services (inpatient or outpatient), financier (financed by social security funds, private insurance by households) and the provider (hospitals classified into 2 types based on

24 The classification of individual consumption by purpose (COICOP) is a classification used to identify the objectives of both individual consumption expenditure and actual individual consumption. The classification of the purposes of non-profit institutions (COPNI) is a classification used to identify the socio-economic objectives of current transactions, capital outlays and acquisition of financial assets by non-profit institutions serving households. The classification of the functions of government (COFOG) is a classification used to identify the socio-economic objectives of current transactions, capital outlays and acquisition of financial assets by general government and its sub-sectors. Non-profit institutions serving households (NPISHs) consist of NPIs which are not predominantly financed and controlled by government and which provide goods or services to households free or at prices that are not economically significant.

25 Statistics Austria "ÖNACE 85 – Health and social work".

whether they are financed by social security funds via DRGs (FKA) or hospitals who contract separately i.e. not via the region, with social security funds[26] (NFKA).

---

26 One notes the stratification by type of provider and by the financing body. This may run counter to a disease-based approach towards output measurement, unless it is understood that the different institutions provide different medical services. The stratification by the financing body is due to the fact that the services provided by social security funds and by private insurance are considered as different products. Both financing bodies offer the same treatments but the private insurance comes with better amenities (reduced waiting times, more comfortable 'hotel' type amenities etc.). Indeed, the fact that consumers are willing to contract private insurance in addition to social insurance indicates that these products should not be matched.

Table 4.4. Overview of country practices in the volume measurement of health services

		Hospital activities		Residential care activities	Medical and dental practice activities		Other human health activities
Country	**Status**	**Acute hospitals**	**Mental health and substance abuse hospitals; Specialised hospitals**		**Doctor services**	**Dental services**	
Austria	Implemented, data since 2001	Deflation with index based on unit costs per treatment by DRGs, cost weights	Deflation with index based on unit costs per treatment by DRGs, cost weights	Number of occupant days, weighted by revenues, no quality adjustment	Number of treatments weighted by revenues, no quality adjustments	64 indices based on fees per single service item paid by the social security, weighted by revenues	Deflation by HCPI
Australia	Implemented	Direct volume index based on DRGs, cost weights	NA	Number of cases by level of care weighted by subsidy rates	Number of services weighted by fees charged	Number of services weighted by cost	NA
Belgium	Implemented in 2009, data since 1995	Direct volume index, based on DRGs, cost weights	Number of occupant days by level of care, weighted by income by category of hospital services	Number of occupant days by level of care, weighted by income by category of hospital services	Number of consultations, use of regulated price of services	Number of consultations, use of regulated price of services	Number of consultations, use of regulated price of services
Canada	Implemented	Deflation with input price index	NA	NA	NA	NA	NA
	Planned	NA	NA	NA	NA	NA	NA
Czech Republic	Implemented	Deflation with index based on sales of hospital per one day care	Deflation with index based on sales of hospital per one day care		Number of treatments	Number of treatments	CPI - component
Denmark	Implemented	Deflation with input price index	NA	NA	NA	NA	NA
	Planned for 2012	Deflation with index based on unit costs per treatment by DRGs, cost weights	Deflation with index based on unit costs per discharge by diagnostic group, cost weights	Deflation with index based on unit cost per patient by type of care, cost weights	Deflation - CPI component	Deflation with index based on unit cost per patient by 2 types of care, cost weights	
Finland	Implemented data since 2000	Volume index based on DRGs, cost weights	Number of day care days	Number of day care days	Number of consultations	Number of consultations	
France	Implemented, data since 1998	Volume index based on DRGs, cost weights	Volume index based on DRGs, cost weights	Volume index based on DRGs, cost weights	Deflation - CPI component	Deflation - CPI component	Deflation - CPI component

		Hospital activities		Residential care activities	Medical and dental practice activities		Other human health activities
Country	**Status**	**Acute hospitals**	**Mental health and substance abuse hospitals; Specialised hospitals**		**Doctor services**	**Dental services**	
Germany	Implemented data since 2006	Deflation with index based on unit costs per inpatient treatment by groups of DRGs, cost weights+ explicit quality adjustment	Number of day care days or number of treatments, cost weights	Number of persons at the end of the year , cost weights by care level	Deflation – unit value for medical/dental services (statutory) and CPI component(private)	Deflation – unit value for medical/dental services (statutory) and CPI component(private)	Deflation - CPI component
Greece	Implemented	Number of day care days	Number of day care days	Number of day care days	Deflation - CPI component	Deflation - CPI component	Deflation - CPI component
Hungary	Implemented, data available since 2001	Volume indices based on DRGs weighted by unit prices s	Volume indices based on DRGs weighted by unit prices	Number of visits	Number of consultations	Number of scores	Number of treatments on basis of services provided
Iceland	Implemented	Deflation with input price index	NA	NA	NA	NA	NA
	Planned	NA	NA	NA	NA	NA	NA
Ireland	Implemented	Deflation with input price index	NA	NA	NA	NA	NA
	Planned	NA	NA	NA	NA	NA	NA
Italy	Implemented data available since 2000	Volume indices based on DRGs, weighted by costs	Volume indices based on DRGs, weighted by costs	Volume indices based on DRGs, weighted by costs	Number of prescripttions	Deflation - CPI component	Deflation - CPI component
Japan	Implemented	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component
Korea	Implemented	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component	Market - CPI component
Luxembourg	Implemented, data since 2000	Deflation - CPI component	Deflation - CPI component	Number of day care days or number of cases by level of care for non market (cost weighted, no quality adjustments); Deflation - CPI component for market	Number of consultations or treatments for non market (cost weighted, no quality adjustments); Deflation - CPI component for market	Number of consultations or treatments for non market (cost weighted, no quality adjustments); Deflation - CPI component for market	Number of consultations or treatments for non market (cost weighted, no quality adjustments); Deflation - CPI component for market

		Hospital activities		Residential care activities	Medical and dental practice activities		Other human health activities
Country	**Status**	**Acute hospitals**	**Mental health and substance abuse hospitals; Specialised hospitals**		**Doctor services**	**Dental services**	
Netherlands	Implemented	Direct volume index based on ICDs by age and discharge numbers + share in day care days as weight	Direct volume indicators based on days of treatments, days of hospitalization and hours of delivered care	Deflation - CPI component (CTG Tariff)	Deflation - CPI component (CTG Tariff)	Deflation - CPI component (CTG Tariff)	
New Zealand	Implemented	Government (non-market) hospitals: Composite volume index based on DRGs, cost weighted; patient discharge and bed-night numbers. Private market: deflation – CPI component	Combined with acute hospitals	Number of employee hours worked	Deflation - CPI component	Deflation - CPI component	Deflation - CPI component
Norway	Implemented	Direct volume index based on DRGs, cost weighted	Number of day care days by levels of care	Number of day care days	Deflation - CPI component	Deflation - CPI component	
Poland	Under construction				Number of consultations	Number of consultations	
Portugal	Implemented	Direct volume index based on DRGs; use of regulated price by DRGs (quasi price)	Direct volume index based on DRGs; use of regulated price by DRGs (quasi price)	Not applicable	Direct volume index based on number of consultations, use of regulated price (quasi price)	Not applicable	Not applicable
Slovak republic	Implemented	Price index derived from	Number		Number of treatments and examinations of adults and children)	Number of treatments and examinations of adults and children)	CPI component

		Hospital activities		Residential care activities	Medical and dental practice activities		Other human health activities
Country	**Status**	**Acute hospitals**	**Mental health and substance abuse hospitals; Specialised hospitals**		**Doctor services**	**Dental services**	
Sweden	Implemented, data since 2003	Direct volume index based on DRGs, cost weights	Direct volume index based on number of days of care by level of care	Direct volume index based on number of days of care by level of care	Direct volume index based on number of consultations, cost weighted	Direct volume index based on number of consultations, cost weighted	Number of consultations or treatments
Switzerland	Implemented	Deflation with input price index	NA	NA	NA	NA	NA
	Planned	NA	NA	NA	NA	NA	NA
United Kingdom	Implemented. Data from 1995. England and Northern Ireland	Direct volume index based on HRGs, cost weights	Direct volume index based on HRGs, cost weights	Proxied by growth in hospital activities (only includes health-related residential care activities)	Direct volume index based on number of consultations, cost weighted	1995-2006 : Direct volume index based on number of treatments, cost weighted. From 2006: proxied by growth in hospital activities.	Proxied by growth in hospital activities
United States	Implemented	Deflation - use of of relevant component of CPI/PPI	Deflation - use of relevant component of CPI/PPI	Deflation - use of relevant component of CPI/PPI	Deflation - use of relevant component of CPI/PPI	Deflation - use of relevant component of CPI/PPI	Deflation - use of relevant component of CPI/PPI
	Planned	Direct volume index based on DRGs, cost weights	NA	NA	NA	NA	NA
Estonia	Implementation behind schedule	Starting to implement DRGs, for the time being use of occupant days by level of care	Starting to implement DRGs, for the time being use of occupant days by level of care	Starting to implement DRGs, for the time being use of occupant days by level of care	Number of treatments	Number of treatments	
Slovenia	Under construction	Direct volume index based on DRGs cost weighted	Number of occupant days by level of care, price weighted	Number of health services provided by type of treatment, price weighted	Number of consultations by type of treatment, cost weighted	Number of health services provided by type of treatment, price weighted	Number of health services provided by type of treatment, price weitghted

4.86. For deflation a number of different price indexes are calculated. The price index for hospital products is based on the average price per DRG – point (= revenues paid for DRG-points divided by the number of DRG points). A DRG point is a scale for representing average costs of a DRG – the higher the costs, the higher the number of points. The price index is calculated as the average price per DRG-point in the current year divided by the average price per DRG in the base year. As the DRGs are reviewed regularly, it is possible to accommodate changes in quality and new products. Small adjustments are made annually and larger ones less frequently at which time new DRGs are implemented.

4.87. The types of hospital services covered by the price index based on DRGs average prices are services to inpatients in general and specialised hospital and hospital based psychiatric services. All services paid by private insurance, (medical fees and non-medical services such as private rooms and food) are deflated by the price index for non-medical services.

4.88. A deflation method is not used for services provided at rehabilitation (or longer stay) hospitals. Production is measured at constant prices by the number of inhabitant days but with the different services grouped according to the main focus of the services e.g. cardiovascular diseases. Nursing services (under medical restrictions) are subject to the same restrictions as rehabilitation services, viz. quantity indicators are available but price per product information is not available. Thus quantity indicators have to be used.

4.89. Outpatient DRGs are under development in Austria. Until they are better developed, the price index for outpatient services is based on the annual average cost per outpatient in hospitals (FKA) weighted by the number of outpatients in the same hospitals.

4.90. The three price indexes for hospital services – the first based on DRG prices, the second used for services paid by private insurance and the third for outpatient services are weighted by their share in production, where the weights are the revenues of the hospitals. The quantity indicator of rehabilitation services is transformed into an implicit price index.

Medical practice services

4.91. Medical practice services are financed by social security, by private insurance or directly by households. Data are available for services financed by social security funds only. Doctors are divided into 18 professional groups (e.g. optometrists, paediatricians). For each of these groups, unit value indexes are produced using the number of cases and the revenues of doctors reimbursed by social security funds. The cases represent the number of treatments and, due to absence of data, are counted by health insurance certificates.

Dental practice services

4.92. Dental practice services are similarly financed by social security funds, by private insurance or directly by households. Available data are those for services financed by social security funds. The price index is based on 3 different types of products for payment. The basic component of the index is the fee for each service item paid for by the social security funds.

Other human health services

4.93. For the remaining item which falls under human health services, the CPI is used for deflation.

Denmark

4.94. To date, Denmark does not apply output-based methods for volume calculations of health services provided by non-market producers. However, Statistics Denmark has put in place a research programme on this matter that is worth reporting on. The present description is based on Deveci, Heurlén and Sorensen (2008).

General description

4.95. In Denmark, the division between market and non-market providers of health services is approximately 17-83 percent, i.e., the vast majority of Danish output is non-market production. The production side of the Danish national accounts records health services activities for hospitals, doctors, dentists, veterinarians and social institutions for adults, in line with the NACE classification of activities.

Hospital activities

4.96. Hospital activities are nearly entirely non-market activities and they account for about 40 percent of all health services. In line with the SNA, the current price value of production of hospital units is measured as the sum of costs. The authors then propose a deflation method to capture volume changes in hospital output. In so doing, they distinguish mainly two products produced by hospital activities, somatic hospital services (COFOG 0731) and specialised hospital services (COFOG 0732).

4.97. The authors construct a (quasi) price index for somatic hospitals on the basis of the Danish National Board of Health's DRG database. As in other countries, the Danish DRG system serves administrative purposes, in particular reimbursement schemes. The DRG system contains information about the number of treatments and the associated fees for around 800 different types of treatment. All data contains information about fees (quasi-prices) and the number of treatments (volumes) for each type of treatment. After excluding non-comparable observations, a Paasche-type price index is constructed as the weighted average of the change in fees per treatment, with the current share of each treatment in total fees as the corresponding weight.

4.98. Psychiatric hospitals are not part of the DRG system, so the method for somatic hospitals was not directly applicable. However, the Danish Register of patients contains details on the number of discharges by diagnostic group and by age. The information on diagnostic groups can be matched with the information on DRG fees by diagnostic group. The data on fees was combined with the data on discharges to construct a (quasi) price index for psychiatric hospitals.

Non-market dental services

4.99. A specific dataset from the Danish Social Resource Statistics provides details on the number of persons receiving treatment. This number is distributed across dental services and orthodontic treatments. The costs for these treatments are registered in another, internal data source on governmental non-market activities (OIMA). However, only total costs are available, with no breakdown into dental and orthodontic services. To deal with this data deficiency, the relative share of each type of treatment was estimated using accounts from two Danish municipalities. Then, unit costs per dental and per orthodontic services were computed and aggregated to yield a quasi price index for non-market dental services.

Residential and day care

4.100. The Social Resource Statistics provide details concerning the number of elderly people who have a place in a nursing home, and the type of care involved. Data from the municipality of Copenhagen was used as the source of unit costs for each type of residence. By combining unit cost information with information on the number of persons in homes, differentiated by type of care, it is again possible to construct a quasi price index.

Germany

General description

4.101. In 2004, the funding system for German hospitals changed from payments of daily rates to a system based on DRGs. The implication is that the price of hospital services moved from an input price to an output price based on diagnoses and treatments. The change in hospital funding triggered statistical research to develop new measures of prices and volumes for hospital services and more generally to review volume measures for the health providing industry. The present description is based on the report (Statistisches Bundesamt 2008) that brings together the findings on volume measurement.

Hospital services

4.102. General hospital services account for the largest part of the DRG system and measurement of their services has been most thoroughly reviewed. Starting with the reporting year 2005, the DRG system provides information on the number of treatments and on the cost weights applicable to each DRG. A major advantage of the DRG-based information is that it covers the universe of hospitals. Because DRG data is the main administrative tool to remunerate hospitals for their services, it can be assumed that the quality of the reported data is high. There are about 1 000 DRGs that are re-grouped into 27 Main Diagnostic Groups. These diagnostic groups form the components of the (quasi) price index that the Federal Statistical Office computes to deflate the value of hospital production (see Box in Section 4.3.1).

4.103. Psychiatric hospitals account for about 4% of total production. There is no DRG system for them, and the charging system has remained with daily rates. Only overall patient numbers are known, there is no information about prices and patient information is not dis-aggregated by severity or type of disease. Deflation is based on the development of unit costs, measured as the treatment costs per psychiatric patient. This corresponds to a volume measure that evolves with the number of patients.

Long-term care and rehabilitation

4.104. These institutions are financed through a system of daily rates and not through a DRG system. Information on the number of cases treated is taken from a different statistical source, a database on the number of diagnoses, broken down by type of disease according to the ICD. However, no corresponding cost information exists. As a proxy for cost weights, the time spent in rehabilitation institutions has been used. A specific drawback of this information is, however, that it comes from a survey that only covers institutions with more than 100 patient beds. This may be problematic if there are small, specialised institutions for which the existing survey is not representative. Other sources would therefore be needed to complement the survey data so as to compute weights for the number of patients.

4.105. An alternative administrative source provides direct information on the number of patient days spent in rehabilitation institutions. The authors carry out sensitivity test to examine the difference in results between un-weighted patient days as output measures and duration-weighted numbers of patients. For the years under consideration, the impact turns out to be small, implying that the average duration of stay per patient has not changed significantly. As the data on un-weighted patient days covers the universe of institutions, the authors recommend using this information for volume measurement.

Quality

4.106. Already in the past, the German Federal Statistical Office has used an explicit quality-adjustment for hospital services measures. A quality index is constructed from four equally-weighted components:

- Average duration of hospital stay;

- Duration of operation;
- Cases per physician;
- Share of therapeutic personnel in overall personnel.

4.107. Between 2003 and 2007, the composite quality index rose by about 1.1% per year. The annual change in the quality index enters the price index additively with a weight of 0.2. Generally, the impact of the explicit quality adjustment on measured prices or volumes is small. Over the years under consideration, the quality adjustment reduced unadjusted price changes for hospital services by between 0.1 and 0.2 percentage points.

Netherlands[27]

General description

4.108. The Netherlands method of calculating an output index for clinical and day treatments is based on data from the Hospital Discharge Register (HDR). In addition, the Netherlands is developing output based methods for compiling output for nursing homes, homes for the elderly, home health care institutes, psychiatric institutes and institutes for disabled patients.

Clinical and day treatments

4.109. Each discharge is counted as a treatment (a discharge also applies to day treatments). The discharges are aggregated into distinct groups in order to calculate volume indices in a meaningful way. For this purpose, the diagnoses were characterised by the 3-digit ICD-9 classification, which resulted in approximately 1000 diagnosis groups. Since age and hospitalisation duration are not independent for most diagnoses, discharges were also subdivided according to age. Data analyses showed that it is useful to group discharges according to 7 age classes (0, 1-14, 15-44, 45-59, 60-69, 70-79, 80 and older). The individual treatments in the HDR are thus grouped into about 7000 diagnosis/age groups.

4.110. The yearly volume changes in output are obtained by calculating a volume index for all the diagnosis/age groups, which are weighted by the fraction of the hospitalisation days in each group on the total number of hospitalisation days over all the groups in the preceding year. Due to a lack of prices on medical operations, the weight of one clinical hospitalisation day is assigned to a day treatment since information is lacking.

4.111. The output method described above adds clinical and day treatments within every diagnosis/age group, as there is a tendency for substituting clinical treatments by day treatments. For example, appendicitis is treated more often with laparoscopy than with traditional surgery. We intend to measure such a shift as a price change and not as a volume change.

Outpatient services

4.112. The HDR covers only clinical and day treatments. Other surveys are used to provide data on other hospital services. Outpatient services are quantified as the number of visits, which in the available

27. Antonio G. Chessa and Foske J. Kleima, (Statistics Netherlands) “The Dutch Experience in Measuring Health Output and Labour Productivity” paper prepared for the joint OECD/ONS/Government of Norway workshop “Measurement of non-market output in education and health” London, Brunei Gallery, October 3 – 5, 2006.

data are not specified with respect to type of specialist. As a consequence, we construct a volume index based on the total number of visits in two successive years.

4.113. Other health services consist of Part-time psychiatric treatments, Rehabilitation, Outpatient births, Haemodialysis and Thrombosis services. The volumes of the remaining five health services are also quantified as the number of treatments or services. Psychiatric treatments and rehabilitation refer to day treatments. The volume of thrombosis services is measured as the number of blood-takings per year.

Other health services, including nursing homes

4.114. Until now, the output volume index numbers for nursing homes and homes for the elderly have been calculated a method consisting of deflating the total production value or budget by the wage-related part of the tariff of a nursing day. Detailed below is a brief investigation of the possibilities for developing at least a B method. This is done by studying the production data for these sectors.

4.115. The main health output of nursing homes consists of nursing, which is expressed as the number of nursing days. It makes up more than 70% of the total budget of the nursing sector. Other outputs are intramural and extramural health services, such as short stays and supplementary care (intramural), household assistance, home care and assistance of elderly patients in their daily activities (extramural). Homes for the elderly are characterised by a similar product composition, with days of care as the main output. Home health care consists of extramural services.

4.116. The product composition of services by institutes for disabled patients and by psychiatric institutes shows only slight variations over time. The main health output consists of treatment days, which are distinguished according to type of handicap for disabled patients (mentally, physically, auditory). Treatments of mentally disabled patients are differentiated further according to severity of handicap, patient's age and institute capacity. Treatments in psychiatric institutes are also differentiated at two levels of detail: by group (addicted patients, children, adults, elderly) and by intensity of treatment or assistance.

4.117. The development of an output volume index for the all health sectors is hampered by the addition of new and changing products in successive years. The extent of this problem is most serious for nursing homes, homes for the elderly and home care, especially from 2003 onwards. Until then, the product composition of these sectors underwent small yearly changes. Fortunately, the main output of each of the five sectors can be tracked from year to year, which makes it possible to develop an output index for these products. The detailed budget and production data should enable the development of a B method for every sector. The treatment of new products and the integration of their output volumes with the output indices of the main products is an open question at this stage.

Norway[28]

General description

4.118. The Norwegian national accounts are introducing direct output methods for non-market health output, with priority given to general and speciality hospital services (inpatients and outpatients, included day care treatments), services from psychiatric hospitals and long term nursing care.

4.119. Due to data unavailability in Norway, the most practical compromise is to use a narrow concept of treatments, which aims at capturing full treatments only within each of the services. This means that if a treatment starts in a hospital, but is finished for instance in a rehabilitation institution, the complete

28 . Brathaug (2006).

treatment is not captured. Only the treatment within each of the institutions is observable. An aspect of the complete treatment issue, which continues to be relevant, is the readmission problem in the case of hospitals or the first visit problem in the case of medical and dental practice services. If a patient has to go back to hospital because of the same illness this means that the original treatment has not yet been completed. A second treatment for the same person is only recorded if the patient is sent back to hospital to be treated for a different disease. A kind of readmission problem also exists for medical and dental practices. In Norway, data regarding readmissions into hospitals are available, but there are data limitations regarding first visits to specialists, general practitioners and dentists.

Hospital services

4.120. Hospital services are non-market as the prices are not economically significant. For all general and specialised hospital services, excluding psychiatric hospitals, a distinction is made between inpatient treatments, which cover both overnight stays and part of day care treatments, and outpatient treatments. For in-patient activities the volume indicator is based on the DRG system adjusted for readmissions. Government owned hospitals and private non-profit hospitals are calculated separately. Both governments owned hospitals and private non-profit hospitals are included in the DRG system and thus, the same method is applied for both categories of hospitals.

4.121. The volume growth for inpatients, outpatients and day cases has been weighted together using cost weights from hospital accounts. From 2002, the hospital accounts are available at a detailed level for all government owned hospitals and all private non-profit hospitals. However, it is a problem that the costs on inpatients and outpatients cannot be separated, so cost weights are based on different information. The assumption is that the costs for outpatients equal the hospitals income related to outpatients. Research has shown that the reimbursement for outpatient treatments from the government plus the out-of-pocket payments from the households (income sources) are too low to cover all costs on treating outpatients, and thus, to achieve a better estimate of the costs the sum of income is multiplied by 1.5.

Psychiatric hospitals/institutions

4.122. The DRG system is not designed for mental hospitals/psychiatric services. The available output indicators are the number of bed days (occupation days), outpatient consultations and day cases, separated for adults and children/adolescents. An indicator based on the number of bed-days is primitive, but presently, this is the only available indicator. The indicator would be acceptable if one could assume that the costs related to a bed-day are equal for all patients and independent of the treatment you receive. Such an assumption seems rather unrealistic. Another problem is the historical cost weights. Even though there is detailed specification of the costs for all psychiatric hospital/institutions, it is not possible to separate costs for inpatients and outpatients up to 2004. From 2005, new functions are included in the hospital accounts, which allow better estimates of the costs related to inpatients, outpatients and day treatments.

Long term nursing care

4.123. Long term nursing care are non-market as the prices are not economically significant. The services can be split between long term nursing homes, old people's homes and combined nursing homes and old people's homes, home nursing and home help. The example given here covers only long term nursing homes. For services from long term nursing homes, occupant days by type of institution (proxy for level of care) can be an acceptable indicator. In Norway, the exact number of occupant days is not available. Only the number of beds available and the number of patients during a year is registered. From these data, the number of occupant days is estimated. Regarding type of institution, it is assumed that the services rendered will be more or less similar for all institutions. Adjusting for quality is problematic. For instance, there has been an increase in the share of single rooms: does this constitute a quality change? If so, how could this be included in the output indicator?

4.124. A new health register covering all individuals who apply and receive nursing and social care services in the municipalities is being established. The data in the register will contain individual information on the person's situation and needs, and in addition information on which (and how much) services are provided. The register will be valuable for statistics about the supply and use of nursing and social care services, and hopefully this register can give better information on output indicators in the future. Until the register is fully developed, the number of occupant days will be used as an output indicator for long term nursing homes.

Services provided by dentists and physicians

4.125. The large majority of services provided by dentists and physicians are market services, and output at constant prices can be derived by deflating with subcategories indices of the consumer price index (CPI). It must be underlined that using the CPI does not solve the problem on how quality changes should be included in the implicit output measure.

Quality

4.126. No explicit quality adjustments are made. Partial quality changes related to product mix are normally captured by a sufficiently detailed product classification. Permanent technical improvements and health research advances make the quality changes in health services an important issue. Adjusting for quality is a challenge. In Norway, good quality indicators to adjust health output are not available. There are problems with a number of approaches. Patient experience surveys are not available, and if they were the surveys will give results that are too subjective to really measure what is wanted. Using outcome indicators we find problematic, as these to a large extent also are based on subjective information from patients. In addition, such surveys are costly and resource demanding to conduct. Receiving outcome indicators every year for national accounts purposes we find rather unrealistic, and basing quality adjustment on indicators that are available only occasionally, will not be a good solution. So before quality adjusted volume indicators are introduced into the national accounts in Norway, further investigation is required.

United Kingdom[29]

General description

4.127. Considerable progress has been made in the way health output is measured in the UK since the first direct measures were published in 1998. Specific recommendations for this approach had been made in the UN System of National Accounts (1993) and subsequently in the European System of Accounts (1995). A significant step in the UK development of methodologies for measuring public sector output was the Atkinson Review (Atkinson 2005). The Review laid out nine broad principles for the measurement of government output and productivity which, together with 54 specific recommendations, form the cornerstone of the work carried out in the UK Centre for the Measurement of Government Activity (UKCeMGA) at the ONS. Estimates of health output and productivity have been published in three Public Service Productivity articles (ONS 2004, 2006, 2008) with the most recent estimate of output growth

29. Lee, Phillip and Little, Christopher E. (UKCeMGA) “Measuring health output and productivity in the UK : an essential element of public accountability” Paper prepared for the joint OECD/ONS/Government of Norway Workshop “Measurement of non-market output in education and health” London, Brunei Gallery, Oct 3-5, 2006.

appearing in a specific paper (ONS 2009). Iterative improvements to the way health output in the public sector is reported in the National Accounts continue to be introduced as a result.

4.128. Growth in NHS output is measured using a cost weighted activity index using base year costs (Laspeyres index). It is calculated and analysed in UKCeMGA before being delivered to National Accounts. The output series omits patients' contributions towards the cost of prescription drugs and dentistry from the cost weights. Currently adjustments for quality change are not included in these compilations, but methodologies that do this are the focus of considerable development, and experimental output measures incorporating such adjustments have now been trialled.

4.129. Volume of activity is measured using various units, depending on the type of treatment or service. There are currently 2500 categories of activities used as building blocks in the calculations of the Department of Health (DH) and Office for National Statistics (ONS), covering around 80% of English NHS expenditures in 2004/2005 (Willmer and Little, 2007). The basis of most activities and cost data used by DH and ONS in their activity index is the English National Schedule Reference Costs (RC). RC collects from the providers the costs incurred to carry out each treatment/service, averages the costs over all providers, and reports the costs per unit of activity measured. This return is mandatory on a yearly basis for all providers of services in the NHS [30].

4.130. To compile estimates for the UK, the ONS has to combine data for England, from the Department of Health, with data from the devolved administrations. Currently national accounts data only reflects data from England and Northern Ireland but data from Wales and Scotland are likely to be incorporated over the next years. The most detailed dataset is for England and currently contains around 15 000 activity categories. Of these around 9 000 are categories of primary care prescribed drugs and 6 000 are categories of hospital and community health services. Most expenditure covered by the index is in secondary care, which comes from a combination of the Hospital Episode Statistics (admitted patient activity data) and the National Reference Cost Schedule (other activity data and all unit costs).

4.131. According to the Atkinson Review, NHS output growth should be adjusted to take account of changes in quality, based on health outcomes directly attributable to the NHS. This is a complex procedure and the data and methodologies necessary to make this adjustment are still being developed. Recent research carried out by the University of York Centre for Health Economics and the National Institute for Economic and Social Research (Dawson et al. 2005), commissioned by the Department of Health, has made it possible to explore the possible impact of some proposed quality adjusters in UKCeMGA work.

Hospital services

4.132. Healthcare Resource Groups (HRGs), the English equivalent to DRGs, are designed to be setting-independent. The same HRGs can be generated from activity within an inpatient, daycase or outpatient setting. There are also specific HRGs to cover Accident and Emergency attendances. There are approximately 1 400 HRGs covering hospital services in England which are categorised within 21 chapters based on body areas.

4.133. HRGs are the main currency used to remunerate acute hospitals for the care they deliver for each admission or attendance under the national system 'Payment by Results'. HRGs have been used for reimbursement, in some form, since 2003-2004. HRG4, the most recent version was introduced for reporting in 2006/07 and is being used for reimbursement from 2009/10. In addition to being used to underpin remuneration of hospitals HRGs are used for benchmarking, planning, commissioning and costing.

30 Hospices and Nursing Homes are excluded from this requirement.

4.134. Approximately 160 HRGs are designed for services that are sometimes, but not always, delivered as part of treatment. These significant costs and activity, such as critical care, chemotherapy, radiotherapy, rehabilitation, diagnostic imaging and renal dialysis generate HRGs in their own right and are unbundled from the core elements of care. This means that an admission or attendance may generate more than one HRG if it includes any unbundled elements. All the HRGs, including the unbundled HRGs are costed seperately as part of the Reference Cost collection.

Other Health Service, including Mental Health

4.135. HRGs are not designed for mental health/psychiatric hospitals. In England, work is currently underway to develop a national currency for adult mental health services provided in all facilities based on an approach of grouping service users into categories based on need. These currencies are based on the nationally mandated Mental Health Minimum Dataset.

4.136. Work has also begun on the development of a national Community minimum dataset that can form the basis of future casemix classifications for community, nursing care and in particular the management of long term conditions.

Quality

4.137. ONS has published estimates of quality adjusted healthcare several times now (ONS 2006, 2008, 2009) but these are not incorporated into UK national accounts yet. The current quality adjustment methodology is based on hospital treatment QALYs, patient experience and primary care outcomes. The QALY method in particular draws heavily on work done for the Department of Health at the Centre for Health Economics (Castelli et al 2008).

ANNEX 4.C: TOOLS TO MEASURE PRODUCTS

Diagnosis related groups (DRGs)

4.138. Diagnosis Related Groups (DRGs) is a system for describing the patient case-mix in hospital care. It was developed by R. Fetter and colleagues at Yale University during the 1970s, initially as a tool for comparing hospital performance to improve cost control in hospitals and more recently has been used for reimbursement of costs (Fetter et al. 1996). The basic idea of DRGs is to describe hospital activity by focusing on the total hospital spell as the final product, measured as discharges defined according to the inpatient's diagnosis and treatment. To classify patients, the International Classification of Diseases (ICD) system is used, in combination with information from statistical studies linking hospital stay to resource use.

4.139. Conceptually, the DRG system assigns patients into groups based on their diagnosis (ICD-codes), procedure codes, age, and the presence of complications and co-morbidities. A key characteristic of the DRG system is that the groups are assumed to be homogenous with respect to clinical and economic resource requirements.

4.140. Since the first application of the DRG system in the United States in 1983, DRG-type grouping systems have been implemented in a number of countries). However, so far no common international DRG system has been developed. This lack of international comparability in the classification is the main difficulty with using DRGs to aggregate hospital volume outputs. This difficulty has been recognised in a number of recent studies (e.g. EU HealthBASKET study and OECD Economics Directorate study on "Analysing Effectiveness in the Health Care Sector"). Different DRG systems have evolved in response to a need to adapt DRGs to local conditions such as differences in clinical practice, strategies for adapting to innovations and demographics.

4.141. Research reported from the EC funded HealthBASKET project found that each of the 9 countries in the study followed their own methodology to determine DRGs. Some countries have moved further away from the DRG system, however. For example, the Netherlands uses a DBC system which is a classification of 29,000 medical procedures[31].

4.142. A move to development of an international DRG system is desirable for ongoing measurement activities in the hospital sector. An international system would also be able to confront differences in the applications of DRGs which limit comparability such as the treatment of outliers and the calculation of cost weights.

4.143. Calculation of cost weights to be used in the reporting of outputs would also have to be standardised for international comparisons. Typically, national DRG systems contain more than 500

31. DBCs are defined as the whole set of activities and interventions of the hospital and medical specialists resulting from the first consultation and diagnosis of the medical specialist in the hospital. One patient, in the same time interval, can have more than one DBC. To describe an episode of care in the DBC case-mix system, at least 3 dimensions have to be specified: the type of care, the diagnosis and the treatment axis. Currently there are approximately 29,000 DBCs as opposed to DRG systems with 600 to 900 classifications. Thus the DBC system is complex with high transition costs. It is still under development (Oostenbrink and Rutten, 2006).

individual groups. Thus, a weighting set is needed when constructing an aggregate output measure for hospital care. Thus, international comparisons of hospitals require the development a common weighting set. Such a set could be built on the basis of a cross-country sample of hospitals which have high-quality patient level data and use cost-accounting. In this regard, a standardised methodology of collecting health expenditure data such as that laid out in the System of Health Accounts becomes of paramount importance. Difficulties arise in the valuation of expenditure on capital and teaching and medical research activities.

4.144. A move to international harmonisation of DRGs however may not be able to solve other problems inherent in DRGs, such as differences in the quality of abstracting data from medical records and variations in the guidelines for coding diagnosis and procedure across countries. Moreover, the assignment of cases to DRG categories may be driven by incentives to maximise reimbursement.

4.145. One further complication in developing a DRG system for international comparisons is the fact that medical procedures and treatments are evolving rapidly. The hospital dataset should thus be designed so as to enable updating.

Outpatient classification systems

4.146. DRGs were initially developed for inpatient services only. There has been some limited development of case-mix based classification systems for outpatient systems as detailed below.

Australia

4.147. The state of Victoria has developed an activity based funding approach for outpatients known as the Victorian Ambulatory Classification and Funding System or VACS. The classification applies to outpatient services of hospitals only. Under the VACS, hospitals are funded on the basis of outpatient encounters where the encounter is defined as the clinic visit, plus all ancillary services (pathology, radiology and pharmacy) provided within the 30 days either side of the clinic visit. The thirty day window has been chosen to encompass the majority of services associated with a particular visit and to enable a reasonable and practical time period for reporting and funding. This approach more closely reflects patterns of clinical care and provides better resource utilisation and controls than the "unbundled" fee-for-service or occasions of service systems.

Canada

4.148. The Canadian Institute of Health Information has developed Day Procedure Groups (DPG) which is a national classification system for ambulatory hospital patients that focuses on the area of day surgery. Patients are assigned to categories according to the principal (most significant) procedure recorded on the discharge database. Patients assigned to the same DPG category represent a homogeneous group with similar clinical episodes and requiring similar resources. Each DPG group is assigned a DPG Resource Intensity Weight value, which is used to standardise the expression of hospital day surgery volumes, recognising that not all day surgery patients require the same health care resources. The volume of day surgery cases is then expressed as total day surgery weighted cases and these weighted cases can be directly compared to the Inpatient weighted cases. There are 115 DPG groups, each one defined by a set of intervention codes from Canadian classification of health interventions (CCI).

Netherlands

4.149. The Netherlands stands alone in its development of the DBC case-mix system which are defined as the whole set of activities and interventions associated with a treatment received in hospitals, outpatient care and/or day care. The ‘diagnosis treatment combinations’ or DBCs were introduced for the registration and reimbursement of hospital and medical specialist care. One patient, in the same time interval, can have more than one DBC. To describe an episode of care in the DBC case-mix system, at least 3 dimensions have to be specified: the type of care, the diagnosis and the treatment axis. The treatment axis includes the setting of either outpatient care, day care or clinical episode. Currently there are approximately 29,000 DBCs as opposed to DRG systems with 600 to 900 classifications. Thus the DBC system is complex with high transition costs. It is still under development.

Sweden

4.150. The Swedish Centre for patient classification systems (CPK) has developed NordDRG-O; a DRG system based on the logic of NordDRG (the common DRG system for Nordic countries) but intended for day surgery and other outpatient procedures like endoscopies. In developing the NordDRG-O, the number of DRGs in the Nordic system was reduced by about one half compared to the NordDRG-System for inpatient care.The groups eliminated were those that do not include a procedure. Most of the groups in the NordDRG-o have the same grouping logic as the corresponding group in NordDRG, but the length of stay must be zero. The system has been used to group Finnish and Swedish outpatient observations with cost data (Fernström, 2002).

United States

4.151. Ambulatory Payment Classifications (APC) have been developed for facilitating prepayment under the Federal Medicare program. Methods for categorising ambulatory case mix are also required in a Health Maintenance Organisation (HMO) setting or where an inclusive capitation rate needs to be set. Ambulatory Payment Classifications classify ambulatory episodes through a grouping process that is based upon ICD diagnosis codes. Services are classified into APCs on the basis of cost and clinical similarity. APCs are described as being to outpatient care what DRGs are to inpatient care, but with a major twist as a single outpatient episode can be assigned multiple APCs.

4.152. From the examples described above, it is clear that there are important differences among countries in terms of the functions covered by the different payment by product systems, and with respect to the identification and measurement of products. A summary of the differences is provided in the Tables below.

Table 4.5. Day care classification systems

Product measurement	Structure
By type	- Same as DRGs for inpatient (England) - Combined with DRGs for inpatient (France; Nordic countries; the Netherlands) - Separated (Canada)

Table 4.6. Outpatient clinics classification systems

Product measurement	Structure	Payment type
By specialty	- Specific list (Victoria, Australia) - Specific list, with separation between "first" and "follow-up" attendance (England)	- Single bundled product
By procedure	- Specific list (England)	- Single bundled product
By type	- Combined with DRG for inpatient (Nordic countries; France; the Netherlands) - Separated (Medicare, United States)	- Single bundled product - Multiple unbundled products

Resource Utilisation Groups (RUGs)

4.153. The Resource Utilisation III model (RUG-III) was developed by Fries et al. (Fries 1994) based on a sample of 7658 Medicare, Medicaid, and private-pay patients in 202 facilities across seven states in the US. The classification system was designed using resident characteristic information and measures of wage-weighted staff time. Staff time measurement studies were conducted in nursing homes in 1990, 1995, and 1997 and were used to establish the 44 RUG-III groups that serve to group patients needing similar levels of clinical resources. RUG-III is a hierarchical classification system with a structure that has forty-four groups. Each group is assigned an index score that represents the amount of nursing time and rehabilitation treatment associated with caring for residents who qualify for that group. The forty-four groups fall into seven major categories: rehabilitation (14 final RUG-III groups), extensive services (3 groups), special care (3 groups), clinically complex (6 groups), impaired cognition (4 groups), behaviour problems (4 groups), and reduced physical function (10 groups). Across most RUG-III categories, patients are divided into a final classification group based on their performance on an index of four Activities of Daily Living: eating, toileting, bed mobility, and transferring. The US Medicare programme uses the RUG classification system to adjust the daily base rates for Skilled Nursing Facilities payment. As from January 2006, a new 53 group RUG classification system replaced the 44 group RUG system. The new system added 9 new payment groups for patients who meet the criteria for "extensive services" and "rehabilitation" groups (MedPAc 2006, 2007). There have been efforts to standardise this instrument across countries. The RUGs system has been adopted in a number of countries including the Netherlands, Japan, Switzerland, the UK, Canada, Spain, Final, Iceland, the Czech Republic,and Italy (certain attempts have been made in Sweden to adopt it) , The output of nursing homes could be aggregated by RUGs.

Population Based Classification Systems

4.154. Several case-mix adjustment measures have been developed to describe the medical problems of a population and their effects on health care consumption/service use during a specified time period (usually 1 year). These measures can be used to allocate resources within health care systems and to risk-adjust patient populations for profiling health services utilisation and providers' performance, measuring the health status, and monitoring outcomes (Majeed, 2001). Health status is the mostly applied dimension of risk and is measured using data from various sources (Rosen, 2003): diagnosis from medical records or claims; self-reported demographics and health status data from patient surveys (e.g. SF-36, Short Form Health Survey); pharmaceutical indicators of patient conditions (e.g. RxRisk). Currently the most commonly used case-mix adjustment measures are diagnosis-based measures, which classify patients into different risk or disease categories based on inpatient or outpatient diagnosis, age and sex gathered from administrative claims/encounter data. These methods include the Adjusted Clinical Group system developed at Johns Hopkins University (Weiner, 1991), the Diagnostic Cost Group developed at Boston University (Ellis, 1995), the Chronic Disability Payment System developed at University of California at San Diego (Kronick, 2000), and the Clinical Risk Groups developed by the National Association of Children's Hospitals and Related Institutions and 3M Health Information Systems (Hughes, 2004). The measurement approaches used by each system are different, but essentially they all work by clustering diagnosis into clinically meaningful categories, which are then aggregated to provide a measure of health status. These categories are used either to explain the type and quantity of resource use and cost at individual level (i.e. concurrent model, same-year resource use) or to predict the future use of health services (i.e. prospective model, following-year resource use). In the context of measuring health volume output, the use of these population-based systems could represent an important step towards the definition of complete treatment: the tools can provide an estimation of services/resource use during a specified time period across all providers and settings by disease.

ANNEX 4.D: QUALITY INDICATORS

Health Care Quality Indicators

4.155. The Health Care Quality Indicators Project (HCQI) (Mattke and Kelley 2006) is developing a set of indicators that can be used to examine health care quality and that can be reliably reported across countries using comparable data. The data set is designed to establish a set of health sector quality indicators that are internationally comparable. Comparative research at the international level has been confined to comparisons of health status indicators such as mortality rates which are measures of overall societal achievement rather than the performance of the health sector. The data set includes both process and outcome indicators as detailed in Table 4.6.

Table 4.7. Areas covered by the current set of OECD health care quality indicators

	Process measures	Outcome measures
Cancer care	Mammography rate Cervical cancer screening rate	Breast cancer relative survival rate Cervical cancer relative survival rate Colorectal cancer relative survival rate
Care for acute conditions		AMI case fatality rate Ischemic stroke case fatality rate Hemorrhagic stroke case fatality rate
Care for chronic conditions	Annual retinal exam for diabetics	Asthma mortality rate Adult asthma hospital admission rate
Communicable diseases care	Coverage for basic childhood vaccination programme Coverage for Influenza vaccination over 65 years old	Incidence of vaccine preventable diseases

QALYs and DALYs

4.156. Two well-known and widely used measures of quantity and quality of life are the DALY (disability adjusted life years) and the QALY (quality adjusted life years). DALYs calculate the loss in terms of years of life in full health, associated with premature mortality and morbidity. Premature mortality is calculated using life tables at birth for males and for females whilst the morbidity weights are calculated by asking a panel of health care providers to assign a value to a set of health states. DALYS measure health outcomes in terms of losses from a normative benchmark. They can be seen as an inverse QALY. QALYs represent levels of quality enjoyed by individuals in particular health starts, while DALYs represent levels of loss of functioning caused by diseases.

4.157. QALYs assign to each period of time a weight ranging from 0 to 1 corresponding to the health-related quality of life during that period where 1 is equivalent to optimal health and 0 is equivalent to death. Negative values are feasible and indicate that some health states are worse than death. The QALY relating to a particular health outcome are then expressed as the value given to a particular health state multiplied by the length of time spent in that state. Generally the amount of time spent in a certain state is approximated by a person's life expectancy. As an example, being on hospital renal dialysis may be assigned a quality adjustment value of 0.8. Thus, if a person spends 20 years on renal dialysis, the QALY is 16. This is assumed to be equivalent to someone living for 16 years in an optimal health state (=1)[32].

4.158. The literature is not conclusive on the reliability of QALYs and DALYs (Dolan, 2000, Gerard and Mooney, 1993). It is, however, undisputed that there are many methodological issues and questions associated with the estimation of QALYs and one of the most problematic issues is how to attach weights to different levels of health related quality of life.

4.159. Moreover, QALYs either do not cover some health condition/treatments or cover them inadequately. Limitations include:

- Less severe health problems;
- Chronic diseases where quality of life is a major issue and survival less of an issue;
- Preventive measures where the benefits may not occur for many years;
- Inadequate weight attached to emotional or mental health problems

4.160. The preferences which determine the value of the QALY are subjective as they are based on individual perceptions of the impact of various conditions on their quality of life. Moreover, different values for the same health state are possible depending on whether the preferences used are those of health professionals, the general public, the patients' families or patients who have experience of the particular medical condition and treatment.

4.161. Overall, it is questionable whether at this stage, QALYs and DALYs can be used for cross-country comparisons, rather than for economic evaluations of health interventions. The value of the measures is in comparing and ranking health interventions within one health system at one point in time. One notes that useful work is going ahead in this area, for instance by Eurostat using EU-SILC micro data.

32 The formulation of health quality status into an index follows two steps. In the first step, questionnaire or interview techniques are used to provide descriptions of a variety of health states or profiles. The preferred measure of describing health states produce a single index score for each state of health which can have a value of 1 or less, where 1 is equivalent to optimal health and 0 is equivalent to death (see Brazier et al., 1999 for more details). In the second step, different techniques can be used to value health state characteristics obtained from questionnaires and convert them into a health utility index. They include the standard gamble, time trade-off, visual analogue scale, magnitude estimation and person trade-off (Brazier et al. 1999). These techniques have been adapted to the valuation of health states from the methods of psychometrics.

CHAPTER 5. MEASURING HEALTH SERVICES ACROSS COUNTRIES

5.1 Introduction

5.1. This chapter of the handbook deals with the comparison of the volume of health services across countries. It describes the approach proposed by a Task Force set up by the OECD to calculate health-specific purchasing power parities (PPPs) with a special focus on PPPs for Hospital services.

5.2. Why do we need health specific PPPs? Health expenditures are probably the most commonly used single indicator of comparative policy analysis in the health sector. They are also of importance in fiscal policy as health expenditure in most countries is publicly funded and represents a large and growing share of governments' budgets. Those seeking to assess health expenditures most commonly benchmark their country's expenditure against international rankings of health expenditure using measures such as health expenditure per capita or health expenditure as a percentage of GDP. While useful indicators for the amount of resources committed, nominal expenditure indicators are sometimes also used to draw direct conclusions about the amount of health care provided. Simple expenditure comparisons, however, cannot take into consideration price and wage differences between countries or differences in productivity between health sectors.

5.3. Health-specific PPPs are meant to address these issues. Health-specific PPPs are ratios of prices (or unit costs) for health services in different countries. Applied to money values of production or consumption expenditure on health for a given year, they yield a volume comparison of health services between the countries under consideration. In principle, PPPs are derived from price ratios of the same products in different countries. In practice, prices are not always meaningful in the health industry and other methods have to be employed to develop PPPs, the spatial deflators. In particular, in calculations of health PPPs, prices are often replaced by unit costs, i.e., by the total costs per unit of medical service provided.

5.2 Temporal and spatial dimension – differences in measurement

5.4. As in volume comparisons over time within a country, volume comparisons at a point in time between countries can be achieved either by directly comparing volumes of health services or by deflating current values with health-specific PPPs. Both approaches require the same steps in measurement in the two dimensions. And they might also use the same sources of information. In this sense, comparisons within a country over time and comparisons between countries at a particular point in time are consistent.

5.5. The main differences between the two dimensions relate to the way products are identified and to the estimation of prices, or unit cost as would typically be the case for health products. Comparisons of volume over time for a given country require within-country consistency of the choice of health products. This means that the product taxonomy has to be stable but it can be country specific. Each country can use its own tools to identify and measure products. Because countries are different, the bundle of products whose quantities or prices are followed will be different for every country. For comparisons across countries, consistency is needed in health product definitions among countries. This means that – in most cases – the country-specific measurement tools cannot be used as they are, but it is necessary to define a common sample of health products.

5.6. The above applies to both the volume and the prices (or unit costs) of products. When PPPs are estimated, ratios in unit costs between countries for a particular health product are weighted with relative importance of each product. To be able to compare unit costs between countries, there has to be consistency of how costs are measured. In particular, the characteristics of allocation methods should be similar. As an example, there are significant variations between countries in which cost items are considered overheads and which cost items can be directly allocated to treatments. Differences in cost classification and scope should be controlled for to avoid biases in spatial comparisons.

5.3 Method used in the OECD-Eurostat PPP comparison programme

5.7. The OECD-Eurostat PPP comparison programme was established in the early 1980s to provide comparable price and volume measures of GDP and its components. As part of this comparison, PPPs for health have been calculated. Expenditure on health appears in three parts in total expenditure on GDP: household consumption expenditure, expenditure of non-profit institutions serving households (NPISHs) and government expenditure on health (see Box).

- Household final consumption expenditure comprises payments made by households for market and non-market goods and services.

- Expenditure on health services provided by NPISHs cannot be identified separately but is included all kinds of expenditure of NPISHs. For most countries, the share of health services provided by NPSHIs is relatively low.

- Government expenditure on health that consists of two main categories:

  - Social benefits in kind, measured as the government reimbursements of pharmaceuticals and health services to households

  - Expenditure of government hospitals, health centres etc. Receipts from sales are subtracted from total expenditure to arrive at net expenditure of government. (Payments made by patients are already included in household consumption expenditure).

Household consumption expenditure

11.06.11.1	Pharmaceutical products
11.06.12.1	Other medical products
11.06.13.1	Therapeutic appliances and equipment
11.06.21.1	Medical Services
11.06.22.1	Dental services
11.06.23.1	Paramedical services
11.06.31.1	Hospital services

Expenditure of NPISHs (include also non-health goods and services)

12.01.11.1	Individual consumption expenditure by NPISHS

Government expenditure on health	
Social benefits in kind	
13.02.11.1	Pharmaceutical products
13.02.11.2	Other medical products
13.02.11.3	Therapeutic appliances and equipment
13.02.12.1	Out-patient medical services
13.02.12.2	Out-patient dental services
13.02.12.3	Out-patient paramedical services
13.02.12.4	Hospital services
Health services provided by government units in kind	
13.02.21.1	Compensation of employees: physicians
13.02.21.2	Compensation of employees: nurses and other medical staff
13.02.21.3	Compensation of employees: non-medical staff
13.02.22.1	Intermediate consumption: pharmaceutical products and other medical goods
13.02.22.2	Intermediate consumption: therapeutical appliances and equipment
13.02.22.3	Intermediate consumption n.e.c.
13.02.23.1	Gross operating surplus
13.02.24.1	Net taxes on production
13.02.25.1	Less receipts from sales

5.8. PPPs for medical goods under *household consumption expenditure* and *social benefits in kind* are based on a "normal" price collection. This means that the estimation of PPP deflators starts by selecting a sample of products in each expenditure category to compare their prices in different countries. Prices to be collected should be average prices for the whole country. A PPP between two countries for a particular expenditure category is in effect a geometric average of all price relatives (parities) formed from a set of product prices belonging to the category. In the OECD-Eurostat comparison programme, countries indicate also whether products they have priced are representative or not. In the averaging, parities based on products that are representative products in both countries, get a higher weight, and parities based on non-representative products (but included in the product list because of their importance for some third countries) are excluded.[33].

5.9. In a multilateral comparison, PPPs are derived for all pairs of countries whenever prices are available in both countries. If all countries price the same products and the products are representative everywhere, geometric averages of the price relatives provide directly transitive PPPs (that is, for countries A, B and C, PPP relatives for A/B and B/C are consistent with A/C) for a basic heading concerned. On the other hand, if a full set of representative prices is not available for all countries, comparisons between pairs of countries will be based on different sets of products resulting in intransitivity. To make results transitive,

[33] This methodological description is strongly simplified. Further information can be found e.g. in the Eurostat-OECD Methodological Manual on Purchasing Power Parities.

a procedure is applied where the final parity for a pair of countries is based on a geometric average of all direct and indirect (via all other countries) parities[34].

5.10. When a country's expenditure is divided by such a PPP, expenditure in two countries is put on an equal basis and can be directly compared.

5.11. It should be noted that prices used in the estimation are "full" prices thus covering payments made by households and government reimbursements. Resulting PPPs are applied as deflators for both household and government expenditure on medical goods when deriving expenditures in comparable terms for a pair of countries. In other words, volume of medical goods is allocated to households and government on the basis of their finance.

5.12. Also expenditure on outpatient services and service prices are subdivided between households and government reimbursements and the same standard methodology has tried to be followed as for medical goods. However, the results have not been satisfactory so far because countries have had difficulties in providing the required price data.

5.13. For inpatient services, an input method is applied for public but also for private hospitals. In the method, PPPs for compensation of employees are based on a wage comparison of employees, that is, wages are used in the PPP estimation in the same way as normal product prices. This means in effect that the productivity of an employee with the same formal qualification is assumed to be the same across countries in spite of sometimes wide differences in the level of economic development. PPPs for other inputs are based on proxy PPPs extracted from other parts of the comparison. As an example, PPPs for pharmaceutical products are used as proxy PPPs for intermediate consumption although the relative difference between prices paid by hospitals and pharmacy prices is not necessarily the same in all countries.

5.14. The purpose of Appendix A is to explain, by way of a more mathematical presentation, the basic approach towards deriving volume comparisons of outputs with the help of PPPs in a non-market context.

5.4 Proposal for the estimation of output-based PPPs for health services

5.4.1 Measurement issues

5.15. The current programme of work of the OECD includes the development of output-based PPPs for health goods and services. The objective is to provide a tool for the comparison of the volume of health expenditure in OECD and EU countries. It also contributes to the broader purpose of deriving economy-wide PPPs for international comparisons of volume GDP.

5.16. There are numerous problems in collecting information that can be used for the development of output-based health-specific PPPs. One such problem arises because the production of many health goods and services are non-market activities. That is, the price of the good or service is not economically significant and cannot be used to represent either the marginal costs of production or the marginal social

---

34 The procedure is the EKS method, named after the three individuals who independently advocated its use: Èltetö, Köves and Szulc. EKS refers to a procedure whereby any set of intransitive binary index numbers are made transitive.
The procedure is independent of the method used to calculate the intransitive binary indices. But, as used in this chapter and in most literature on the subject, EKS covers both the way the intransitive binary PPPs are calculated and the procedure to make them transitive and multilateral.

value[35]. This may also be true of health goods and services which are provided by market producers because many health expenditures are subsidised by social insurance. Thus, reliable information on prices of health goods and services is often very difficult to obtain and often not available.

5.17. Aside from the lack of significant price information, the complexity and variety of health goods and services means that it is often difficult to ensure that the same goods and services are being compared across countries. This problem was evident in the EU HealthBASKET project where there is a problem with comparison of DRGs because the mix of interventions which makes up a DRG can vary. Mechanisms for remuneration of general practitioners (or family physicians) across countries can vary and may be based on salary, capitation or fee-for-service. The different remuneration patterns create different incentives so that the service received from a salaried doctor may be quite different to that received from a doctor who raises a fee for each service rendered. Thus, institutional differences in the organisation of health services potentially lead to different prices (where they exist) but also differences in both the quantity and quality of the service received.

5.18. This chapter focuses on hospital services because of the large share of total health costs that are consumed by hospitals and the measurement difficulties outlined above. In particular, market prices are in general more available for health products other than hospital services.

5.19. As a starting point, the proposed approach relies on comparing hospitals in terms of the volumes and type of activities they produce without explicit quality adjustments. This means that the same well-specified health service is assumed to be delivered with the same level of quality. That is, "one is (also) implicitly assuming that there is no difference between organisations in the effectiveness with which they implement the procedures" (Jacobs, 2006: 27).

5.20. One of the major consequences of the absence of markets for hospital services is that there are no prices to reveal patients' marginal valuations of health care outputs. Thus, in line with the literature (e;g., Castelli, 2007; Triplett, 2003) it is proposed to use costs to value output.

5.21. An important decision in the study design relates to how specific the description of output (i.e. cost object) should be for the hospital products to be comparable across countries. In order to identify, measure, and value products, three options could be used, each involving different strengths and weaknesses:

- *Per patient*. A case of hospitalisation is the cost object. A profile of care and a profile of costs is estimated "bottom-up" at patient level. A similar approach was proposed and used in the HealthBasket project funded by the European Union (Schreyogg, 2005), and proved to be feasible for 10 common care episodes (including five in the outpatient and five in the inpatient setting) from 9 countries. However, the approach presents a high variability of unit costs per case among countries; it is based on standard cost, which often needs an ad hoc detailed data collection; and it is difficult to ensure that the data are representative within and across countries. More important, "micro-costing requires substantial resources, the amount of which may exceed the benefit of this approach" (Schreyogg, 2008).

- *Per diagnosis or procedure category*. The output is defined as simple aggregations of cases that have been coded: each inpatient case is assigned to a category on the basis of a list of codes that correspond to the disease or intervention. An example is provided in the Hospital Data Project funded by the European Union (Magee, 2003). The project aims to maximise the

35. A price which is not economically significant is deliberately fixed well below the equilibrium price that would clear the market.

statistical comparability of hospital activities, using data routinely collected by countries and mapping tables from local procedure coding schemata to ICD-9-CM codes. A major limitation is that the project focuses only on the product identification and measurement phases, and the match of the product categories with the costs incurred to provide those services is not within its scope.

– *Per group* (e.g. Diagnosis Related Groups). The DRG system represents a classification of hospital activity (i.e. case-mix) based on relatively homogeneous cases. Each inpatient case is assigned to a DRG on the basis of the diagnoses, procedures, age, and discharge disposition information available in the minimum basic data set for acute inpatient care. For inpatient hospital services (Berndt, 2000:143-144) "DRGs represent the beginning of a structure which could facilitate defining, measuring, and pricing the output of medical care providers…in particular (they) involve the treatment for an episode of hospitalization for a particular condition/diagnosis. Instead of pricing each of the components of a hospitalization, with DRGs the complete bundle of hospital services is given a single price". They are currently used as the pricing unit for inpatient hospital services in the estimation of the Producer Price Index and the CPI by the Bureau of Labor Statistics (Cardenas, 1996). DRG unit costs and cost weights (i.e. relativities in terms of resource consumption) are also determined from a series of specific cost finding studies and used as a basis to set prices at national level. These studies are undertaken mostly by a sample of hospitals every two years using specific costing guidelines (Department of Health of the U.K., 2008; Australian Government, 2006).

5.22. In time series comparisons, within country consistency of measurement is necessary. Thus country specific taxonomies, such as DRG systems, may be used. In cross country comparisons, however the product descriptions must be as consistent as possible. As most DRG systems tend to have country specific modifications, using DRGs for comparison purposes is problematic.

5.23. In light of testing the feasibility of applying the methodology at country level, the OECD has designed a qualitative questionnaire to verify data availability at hospital level to feed into the proposed methodology and to identify differences which may be controlled for in the PPPs estimation process. Annex B describes the results of the questionnaire that was filled out by OECD member states, and non OECD-EU countries in 2008.

5.24. Despite the widespread use and availability of case-mix measurement (European Hospital and Healthcare Federation, 2006; Roger France, 2003), there are a number of problems with using DRGs.

- There is no international DRG system although there is a connection between many of the classifications which make partial harmonisation possible[36]. There has been some work done on comparability of DRG systems (see for example Schreyogg et al., 2006);

36 A necessary first step for in the development of an international DRG system is the development of an International Classification of Health Interventions (which could then be used to group different type of treatments in some more concise international DRG system). The development of an International Classification of Health Interventions (ICHI) was discussed at the 2006 meeting of the WHO Family of International Classifications (WHOFIC). Given the feedback received on the beta field tests of an early version of the ICHI (which was based on a simplified version of the Australian Classification of Health Interventions), it was recognised that the overall construction of an ICHI needs to be revisited and there is a need to start to develop a new ICHI. This developmental work is expected to be carried out by a Work Group reporting to WHOFIC, and it can be expected to take several years (at least 5 years) to come up with such a new international classification system. To illustrate the magnitude and complexity of the task, the Canadian Classification of Health Interventions contains no less than 18,000 different codes and the numbers are similar in other classifications such as the French Classification Commune des Actes Médicaux (CCAM).

- Some OECD and EU countries have not implemented a DRG system;
- Some countries have applied the DRG system to a limited number of hospitals;
- Not all inpatient services are included. For example, they have not been designed for psychiatric hospitals and institutions and other types of chronic or long-term care.

5.25. This chapter adopts an approach that uses the strengths of options two and three described above. It uses diagnosis and procedure categories to identify products, secondary data sets containing patient level coded information to measure products, and official unit costs data bases to attach a value to products. That is, it identifies products in terms of case types (i.e. categories) for which unit costs are available through national cost studies carried out to calculating payment rates for DRGs[37].

5.4.2 Steps in measurement

5.26. The steps in the measurement of health products for PPPs estimation are: (a) identification of homogeneous products, (b) measurement of quantities of products, (c) placing a value on products.

Identification of homogeneous products

5.27. In the PPP framework, the items for which costs are collected and reported should be comparable and representative within a basic heading (Eurostat-OECD, 2006). It is clear that such a list of items will not be exhaustive in that it covers all the activities within a type (as an example, inpatient hospital care). However, exhaustiveness is not necessarily required if the selected case types are considered representative for a broad set of activities.

5.28. Products are said to be comparable if they have identical or equivalent physical and economic characteristics – that is, if they have the same or similar technical parameters and price determining properties. In the context of PPPs, equivalence or similarity between products is defined as meeting the same needs with equal efficiency so that purchasers are indifferent between them and are not prepared to pay more for one than for the other. For non-market hospital products, this requirement is not easily fulfilled. For this reason, it may be necessary to rely on the identification of products that have the same or similar technical parameters, that is, they are homogeneous in terms of expected treatment profiles.

5.29. Representativity is a concept that relates to individual products within a basic heading. A product is called representative if it is generally found in the market and expenditure on the product accounts for (although not necessarily) a significant proportion of the total expenditure of all products covered by the basic heading. Normally, representativity of products cannot be accurately determined due to lack of information on expenditure proportions. In the context of non-market hospital products, the representativity requirement would be met if the case types represent common occurrences of treatments that are typically dealt with in hospitals and/or if there is evidence that the total costs of the selected products account for a sizable share of total hospital costs[38].

5.30. The proposal foresees the categorization of case types into medical and surgical, the former being those cases for which no operating room procedure was reported to be performed. The following are the

---

37 In the remaining part of the document, we use the term DRGs to identify the case-mix measurement tool in general terms. We recognize that the taxonomy of the classification system varies across countries and that different systems are in place, such as Groupe Homegène de Malade in France or Australian Refined DRGs in Australia.

38 In line with current practice in PPP calculations, no quantified minimum percentage is needed here.

criteria that could be used for identifying the hospitals case type list in the PPPs context. The case types should:

- Represent common procedures or diagnoses;
- Account for a significant percentage of hospital expenditures;
- Identify cases likely to be admitted to a hospital;
- Represent procedures which are elective and are the principal procedure within one hospitalisation (for surgical case types);
- Represent well identified conditions without complications and/or comorbidities (for medical case types).

5.31. Annex C contains a proposed list of case types. The list is based on a proposed shortlist of inpatient case vignettes presented during the 1st Task Force[39] meeting (Huber, 2007), on the proposal by the Expert group on procedures under the Hospital Data Project (Smedby, 2007), on the list which is currently used at the OECD for Health Data collection (OECD, 2007), and on discussions held during the Task Force meetings. The criteria listed above were used for the selection of case types.

Measurement of quantities of these products

5.32. The comparison uses routinely collected information on inpatient activity and costs as a basis for data gathering and collating. The proposed approach limits the costs of collecting data by using secondary databases. Moreover, it may have the advantages of larger sample size and greater data validity. On the other hand, this approach requires an analysis of the classification logic used in each country to identify and measure case-mix, given the important differences among countries in the grouping logic.

5.33. To measure those case types, secondary data sets containing coded diagnosis and procedures information will be used. The first step of the process envisages the selection of International Classification of Disease, ninth revision, clinical modifications (ICD-9-CM) codes that identify each case type (column 2 of the tables contained in the Annex C) and the specification of rules for the selection of cases (column 3).

5.34. To handle the variations of coding schemata and ensure comparability of products definition, mapping tables have to be used from local codes to ICD-9-CM. For medical case types, the WHO International Shortlist for Hospital Morbidity Tabulation (ISHMT) (WHO, 2006) contains a mapping of ICD-10 diagnoses codes to ICD-9 diagnosis codes. For surgical cases, mapping tables for selected country

39 The Task Force for the Development of Health-Specific Purchasing Power Parities (TF) is a technical group whose roles and responsibilities include: development of concepts and methodology; assessment of data availability and feasibility of data development; formulation of joint OECD/Eurostat guidelines for data collection for the purpose of health PPPs. All member countries of the OECD and non-OECD European Union members were invited to nominate participants to serve on the TF. Members of the TF are invited to participate actively in the programme of work by reviewing and commenting on papers and issues at TF meetings and preparation of papers on topics of interest. The leadership and administration of the TF is with the OECD. The TF reports in the first instance to the OECD Health Accounts Experts Group, and also to the OECD Health Committee and the OECD National Accounts Working Party. The reports of the TF are also made available to the Eurostat PPP Working Group. The TF sends its reports also to the Eurostat Partnership in Health Technical Group Care and the Eurostat Working Group on Public Health Statistics. The first meeting of the TF took place in Paris on 8 June 2007; the second one on 7-8 February 2008.

procedure classification systems to ICD-9-CM are available through the Expert group on procedures under the Hospital Data Project.

5.35. The correspondence between each of the case types and a DRG is then reviewed through a qualitative evaluation of the taxonomy of the classification systems used by the individual country. The review of the contents of DRGs Definition manuals will allow us to state if:

- The case type corresponds to a DRG;
- The case type corresponds to more than one DRG. This is the case when the taxonomy uses factors as age or severity to split products corresponding to one case type;
- The case type is included in one DRG, and there are other case types included in the same DRG. This is the case when the clinical and resource consumption analysis does not justify the representation of a case type by a DRG (or family of DRGs).

5.36. For surgical case types, the review of the DRGs taxonomy will allow us also to evaluate if there are significant differences among countries as to the diagnoses for which the specific type of procedure is expected to be performed.

Placing a value on products

5.37. A universally accepted costing methodology does not currently exist in the health care sector. Approaches to resource consumption measurement vary widely, and are determined by the objective of the cost analysis and the availability of data (Mogyorosy, 2005; Wiley, 1993). On one end of the spectrum, there is the gross-costing or top down approach. On the other end, there is the bottom-up or micro-costing approach.

5.38. The gross-costing approach is essentially a product line (or case-mix) cost accounting model, with the core objective of costing individual patients grouped into similar classes. The basic information comes from the hospital's general ledger. The reference Yale cost model[40] has four basic steps (Fetter, 1976):

- Definition of the initial cost centres of the hospital, which include overhead cost centres or support services, ancillary services, and wards;
- Allocation of overhead costs to the remaining cost centres, ancillary services, and care units;
- Allocation of ancillary services to the wards. This may be done according to actual use of ancillary services or to indicated use;
- Allocation of the final cost centres to patient level/DRG groups, according to actual use of resources or some allocation statistics.
- Identification of clinical activities to describe treatment profiles;
- Identification of the discrete resource areas/intermediate cost centres which provide the identified activities;

40 The so called Yale cost model was developed by the Health System Management Group established in 1977 at Yale University. This entity, with Professor Fetter as director, has carried out all DRG-related work at Yale.

5.39. In the direct identification and measurement of patient-specific service delivery process, frequently called bottom-up or micro-costing approach, five steps are taken in the process of defining patients' resource use and cost profiles:

- Definition of appropriate workload measures for each resource area;
- Identification of individual patients' treatment profiles, in terms of workload units of the relevant clinical activities performed;
- Place a value on workload units from each resource area in order to derive individual patients' cost profiles.

5.40. Several countries measure resource utilisation retrospectively through a top-down or mixed approach. The hospital costing process usually begins with the general ledger. The purpose is that a control total for costing is established, and this should be the full cost of providing services to patients. The cost object could be either the DRG, or the case type, or the patient. The main difference relates to the allocation of costs from ancillary cost centres and direct cost centres to case types or patients. The most common methods of allocation are cost-to-charge ratios, weighted length of stay, actual costs, relative value units, and standard costs (Chandler, 1991). As an example, the English NHS requires a standard clinical and resource profile to be determined for the HRGs that cover at least 80 % of cost and activity at each point of delivery within a hospital. In Australia, there are service weights by AR-DRG available for each ancillary service. The results of these studies are benchmark or reference costs, and usually reflect the average observed cost by product. Costs are calculated on a full absorption basis to identify the full cost of services delivered.

5.41. The analysis of the correspondence between case types and DRGs is the basis for attaching a value to products. This will be done using the official unit costs values that are available through national cost studies. With respect to the three options described above, a unit cost will be computed as follows:

- The case type corresponds to a DRG. This is the easiest one, because unit costs values can be used as they are published;
- The case type corresponds to more than one DRG. In that case a weighted average of the unit costs by DRG can be used, the weights being the number of cases in the national sample. The case type is included in one DRG, and there are other diagnosis or procedures included in the same DRG.

It is proposed that countries will be asked to control for the within-DRG variability. This might be feasible provided that countries use allocation statistics at patient or diagnosis/procedure type or day of stay level.

5.4.3 Quality adjustment

5.42. The PPP proposal provides a first measure of quality through the differentiation of services by case type. This is one of the approaches to the measurement of quality, proposed for instance in the United Kingdom (Atkinson, 2005). The idea is that products of different quality are treated as different products. The validity of this assumption may have to be reassessed, particularly when sufficiently homogeneous service products cannot be identified in all countries.

5.43. Other ways towards capturing quality include direct measurement of the marginal contribution of health care activity to health outcomes and application of explicit quality adjustment factors to the PPPs as described above. They are discussed in the context of comparisons over time in Chapter 4. Even in a within-country, temporal comparison, explicit quality adjustments are difficult, conceptually and empirically and at this point, this avenue has not been pursued for cross-country comparisons where problems are likely to be even more important.

ANNEX 5.A: INPUT AND OUTPUT-BASED PPPS

5.44. The purpose of this appendix is to explain, by way of a more mathematical presentation, the basic approach towards deriving volume comparisons of outputs with the help of PPPs in a non-market context.

Background – a market situation

5.45. It is easiest to start by considering a market situation with a well-defined product, say a common type of apple[41]. The price of an apple in country A, P_A, can be observed and compared to the price of the same type of apple in country B, P_B. The purchasing power parity for apples between country A and country B is simply the ratio between these two prices:

(1) $PPP \equiv P_A/P_B$

5.46. Along with prices, there is information on the total expenditure E on apples, which is the product of prices and the number of apples Y consumed in each country:

(2) $E_A=P_AY_A;E_B=P_BY_B$

5.47. The real value or volume index of apples between the two countries is calculated by deflating relative expenditure with relative prices, i.e., PPPs. In our simple one-product case this obviously yields the quantity relative:

(3) Volume index of output = $(E_A/E_B)/\ PPP=Y_A/Y_B$

5.48. With more than one product involved, a weighted average of different products has to be constructed (see discussion below) but the principle of deflation is the same as for the single product case at hand. So far, nothing has been said about outputs and inputs in production processes but apples are of course produced goods and the 'P's are the market prices of the output of apple production. What about inputs? Take a very simple case where there is only one input into growing apples, labour, which we shall measure by the number of hours worked, L. Employees are paid the wage rate w, so that the total costs of apple production in the two countries are

(4) $C_A=w_AL_A$; and $C_B=w_BL_B$.

41 PPP specialists and agricultural experts would object that there are in fact many types of apples, and the same type of apple may be representative for apple consumption in one country but not in the other. For the purpose at hand, we ignore this point although the question of comparability and representativeness carries of course over to the non-market case and to health services.

5.49. When competition is intense, all extra profits will be competed away, and in this simple case, the revenues of apple farmers will just cover their costs. In a more complex setting, profits and capital remuneration could be introduced but this is not needed for the basic point to be made here. So, if in a competitive situation, revenues just equal costs, we have:

(5) $E_A=C_A$ and therefore $P_AY_A=w_AL_A$ as well as

$E_B=C_B$ and therefore $P_BY_B=w_BL_B$.

5.50. One other element is needed before we shall examine non-market production, namely an indication of production technology. Note that all indications above were nominal or accounting relationships but we are also interested in how labour input is transformed into outputs of apples. Say that there is a simple linear input-output technology in each country that links labour and apple harvest:

(6) $Y_A=\alpha_AL_A$ and $Y_B=\alpha_BL_B$.

5.51. A volume index of inputs is the ratio of the two countries' labour input:

(7) Volume index of inputs = L_A/L_B

5.52. Alternatively, we could have derived the volume index of inputs by deflating wage costs with the salary deflator, or 'wage PPP' w_A/w_B:

(8) Volume index of inputs = deflated costs = $[C_A/C_B]/[w_A/w_B] = L_A/L_B$

5.53. And by combining (6) and (3) it is apparent that the volume index of outputs equals the volume index of inputs, adjusted by the relative productivity α_A/α_B:

(9) Volume index of output = $Y_A/Y_B = [\alpha_A/\alpha_B]\ [L_A/L_B]$

Output-based PPPs – a non-market situation with a single input

5.54. We are finally ready to look at non-market production. Suppose that apples are produced only by a single publicly-run farm and sold at a price that has been administratively determined and may well be below cost or even zero. In such a case, it is not helpful to measure the money value of apple farming by observing sales revenues because they are driven by the administrative price and may even be zero if for some reason the government gives away apples for free. In a non-market situation it makes sense to measure the money value of output by the sum of costs, so the value of apple production in our two countries would equal each country's wage bills, C_A and C_B. The assumption that the value of output equals the value of inputs is the standard practice for non-market production in national accounting.

5.55. How should the volume or quantity of apple production be compared between the two countries in such a case? There are two possibilities.

Input-based PPPs and measures of production

5.56. The first option is to assume that productivity between the two countries is the same, i.e., $\alpha_A=\alpha_B$. In this case, the volume index of output can readily been derived from simply comparing labour inputs, by using (9):

(10) Input-based volume index of outputs = $Y_A/Y_B = [\alpha/\alpha]\ [L_A/L_B] = [L_A/L_B]$

5.57. Rather than comparing the volume of inputs directly, we could also have measured the 'wage PPP' w_A/w_B in the two countries first and then applied them to deflate each country's costs as in (8). The result would have been the same, namely a volume measure of inputs that is used to approximate the volume measure of output by assuming that there are no productivity differences between countries.

5.58. This is an unsatisfactory situation because measuring productivity and performance is important, but to do so, one needs volume measures of output that can be compared to volume measures of input. When the volume of output is set to be in fixed proportion to the volume of inputs, this is not possible. Hence the quest for output-based PPPs and volume measures of production.

Output-based PPPs and measures of production

5.59. In our apple example, then, the second option towards measuring a volume index of output is to directly count the number of apples in each country and compare Y_A and Y_B. In principle, this is also an option for more complex situations, for example for health care, but from a practical viewpoint it is often easier to derive measures of production with the help of output-based PPPs.

5.60. This is the alternative but conceptually equivalent way of directly counting apples, akin to the comparison of market prices in (1) that were subsequently used for deflation. In the non-market case, there is one important difference: the market prices cannot be directly observed and now have to be replaced by *estimated prices* p^e, and the simplest way to define these quasi prices is as the unit cost of an apple, emulating the competitive situation where the market price just equalled average costs, i.e., total costs divided by the number of apples:

(11) $p_A^e = C_A/Y_A \equiv c_A$ and $p_B^e = C_B/Y_B \equiv c_B$

Our *estimated output PPPs* are then

(12) $PPP^e_{A/B} = p_A^e / p_B^e = c_A/c_B$

5.61. Just as in the market case, we can use these PPPs to deflate the money value of output, keeping in mind that the money value of output was set to equal the nominal costs. Deflation yields:

(13) Deflated value of output $= [C_A/ C_B]/ PPP^e$

$= [C_A/ C_B]/ [p_A^e/ p_B^e]$ using (12)

$= Y_A/Y_B$ using (11)

5.62. From (13), it is clear that direct count of apples yields the same result as deflating the money value of apple production by estimated PPPs. Arriving at (13) may look like a bit of superfluous exercise because in order to get to (13), we need to estimate unit costs and to do so, we had to divide total costs by the number of apples – so why not count the number of apples in the first place? The reason is that in practice, it is often easier to measure unit costs c_A or c_B directly, just as it is typically easier to compare prices in a market situation and use them to deflate the money value of output instead of directly comparing quantities. Furthermore, if not all prices or all quantities are measured, it is more reasonable to assume that a particular price P^e is representative for a broader number of goods than a particular quantity Y.

5.63. But what exactly marks the difference between the output-based deflation procedure here and the input-based deflation procedure discussed earlier? As there is only one input in our example, labour, the unit cost measures c_A and c_B would seem to be very close to measures of relative wages, w_A/w_B that were used for deflation in (8). But they are not and in fact this distinction distinguishes input and output-based

PPPs and volume measures of production. The input-based or wage PPPs represent the costs of inputs *per unit of input*, whereas the output-based PPPs represent the costs of inputs *per unit of output*. For apple production, the former corresponds to the wages per hour worked, whereas the second corresponds to wages per apple harvested. It is quite straightforward to show that these two measures differ by the level of labour productivity, because

$c_A = w_A L_A / Y_A = w_A / \alpha_A$ and $c_B = w_B L_B / Y_B = w_B / \alpha_B$ and therefore

$[c_A / c_B] = [w_A / w_B] / [\alpha_A / \alpha_B]$

5.64. As a first conclusion we can say that in a non-market context, even when the money value of output is measured by the money value of inputs, it is possible to derive the right volume indices of output Y_A/Y_B, as long as the estimated PPPs are based on cost measures per unit of outputs, and not on cost measures per unit of inputs. This conclusion carries over to the multi-product case. The volume index of output will then be a weighted average of volume indices of different products. Each product's share in the total money value of outputs (equivalent to each product's share in total costs of producing products) constitutes the weight to be used[42].

5.65. This is not the end of the story, however, because we have to add an additional element of reality into our considerations. It was said earlier that it may often be easier to collect unit costs rather than direct quantity measures. But unit costs, and this is where our considerations become more realistic, are typically themselves a set of several cost components. In other words, the assumption that only one homogenous type of labour is needed to produce apples needs to be relaxed. In so doing, we shall move to health services, the actual object of interest.

Output-based PPPs – a non-market situation with a several inputs

5.66. Instead of apples, let us turn our attention to health care and consider as an example the treatment of a stroke. We shall say that there are two types of inputs needed for this medical service: physicians' time (L) and drugs (D). And as would be the case in practice, an international, output-based comparison would start out by looking at the *unit costs per treatment of a stroke* in various countries. For each country, the costs of treatment of a typical stroke would be summed up from labour costs and costs for drugs:

(14) $c_A = [w_A L_A + v_A D_A] / Y_A$ and $c_B = [w_B L_B + v_B D_B] / Y_B$,

where v_A and v_B is the price per drug in country A and country B, respectively. An interesting question arises here: to set up our estimated output PPPs, PPP^e, should we proceed by type of unit cost and then combine them to arrive at PPP^e or should we use the total unit cost c directly? The first avenue was followed by Schreyögg et al. (2008) in a recent computation of international health PPPs, so it is worth examining its implications.

42 See Diewert (2008) for a more general discussion.

5.67. Let us explore the second option first, where estimated PPPs are directly computed as the ratio of total unit costs in two countries. In this case, the estimated output PPPs are given by

(15) $PPP^{e} = p_A^{e} / p_B^{e} = c_A/c_B$

where c_A and c_B are the multiple-product unit costs as defined in (14). Deflating total costs for each country by this price ratio yields:

(16) $$\begin{aligned}\text{Deflated value of output} &= [C_A / C_B] / PPP^{e} \\ &= [c_AY_A / c_AY_A] / [c_A / c_B] \quad \text{using (15)} \\ &= Y_A/Y_B\end{aligned}$$

5.68. Thus, the second option is a valid way towards deriving volume ratios of the medical service because deflation with output-based PPPs is again equivalent to direct output measurement.

5.69. Turning to the first option where several PPPs are constructed, one for each subcomponent of the total unit cost. The estimated PPP for physician's time ($PPP^{e,L}$) and the estimated PPP for drugs ($PPP^{e,D}$) are

(17a) $$PPP^{e,L} \equiv p_A^{e,L} / p_B^{e,L} = [w_AL_A/Y_A]/[w_BL_B/Y_B]$$

(17b) $$PPP^{e,D} \equiv p_A^{e,D} / p_B^{e,D} = [v_AD_A/Y_A]/[v_BD_B/Y_B]$$

5.70. The next step, following Schreyögg et al. (2008) is to form a weighted average of these individual PPPs, based on the cost shares of wages and drugs in production. While the cost shares of labour and drugs are natural candidates to weight the individual PPPs, an immediate question arises, namely whether country A's or country B's cost structure should be applied. What is the economic intuition behind such a multi-step aggregation method? The basic assumption behind aggregating with one country's weight is that the so constructed PPPs control for the mix of treatments between countries that is reflected in the cost.

5.71. Thus, if country A has a treatment of strokes that is much more based on drugs than on physicians' intervention than country B, and if this difference has an effect on total costs per treatment of stroke, such a cost difference would show up in total PPPs and therefore in volume output measures and productivity comparisons, if the total unit cost method (15) is applied. In the stepwise unit cost method, any differences in total costs that are due to a different mix in treatment methods, would not show up in PPPs and therefore yield a different volume and productivity comparison.

5.72. We conclude that the basic conclusions arrived at in the context of the apple production carries over to the more complex case with multiple cost categories: PPPs are output-based, if the underlying cost comparison is per unit of output, and not per unit of input. There are different ways of aggregating across cost categories, depending whether or not it is desirable to control for differences in the cost structures between countries. At first sight, it would not appear desirable to control for such differences. Substitution between types of treatments or between inputs can be a source of differences in efficiency, and it may be desirable to identify such differences in productivity comparisons.

ANNEX 5.B: AVAILABILITY OF INFORMATION IN NATIONAL HOSPITAL DATASETS

Country	Diagnosis classification system	Procedures classification system	DRG system
Australia	ICD-10-AM	ICD-10-AM	AR-DRG
Austria	ICD-10	Austrian Catalogue of Procedures	Austrian DRG
Belgium	ICD-9-CM	ICD-9-CM	APR-DRG
Bulgaria	ICD-10 (3-digit level)	-	-
Canada	ICD-10-CA	CCI	CMG +
Czech republic	ICD-10	National nomenclature	IR-DRG
Denmark	ICD-10	NSCP-D	Dk-DRG
Estonia	ICD-10	NSCP	Nord-DRG
Finland	ICD-10	NSCP-F	Nord-DRG
France	ICD-10	CCAM	GHM
Germany	ICD-10-GM	OPS	G-DRG
Greece			
Hungary			
Iceland			
Ireland			
Italy	ICD-9-CM	ICD-9-CM	CMS-DRG
Japan	ICD-10	Local system	DPC
Korea	ICD-10	Local system	K-DRG
Latvia	ICD-10	National nomenclature	L-DRG
Lithuania	ICD-10	National nomenclature	-
Luxembourg	ICD-10 (3-digit level)	National nomenclature	-
Malta	ICD-10	ICD-9-CM	-
Mexico			
Netherlands	ICD-10	National nomenclature	DBC
New Zealand	ICD-10-AM	ICD-10-AM	AR-DRG
Norway	ICD-10	NSCP-N	Nord-DRG
Poland	ICD-10	ICD-9-CM	-
Portugal	ICD-10	ICD-10	AP-DRG
Romania	ICD-10-AM	ICD-10-AM	AR-DRG
Slovak republic			
Slovenia	ICD-10-AM	ICD-10-AM	AR-DRG
Spain	ICD-9-CM	ICD-9-CM	AP-DRG
Sweden	ICD-10	NSCP-S	Nord-DRG
Switzerland	ICD-10	ICD-9-CM	AP-DRG
Turkey	ICD-10	National nomenclature	-
United Kingdom	ICD-10	OPCS	HRG
United States (Medicare)	ICD-9-CM	ICD-9-CM	MS-DRG 25.0

ANNEX 5.C: LIST OF CASE TYPES

ICD-9-CM codes and rules by inpatient medical (IM) case types

Case type number	*Case type description*	*ICD-9-CM codes (version 25 FY 2008)*	*Rules*
IM01	Acute myocardial infarction	410.00-410.92	No operating room procedure is performed.
IM02	Angina pectoris	413.0; 413.1; 413.9	No operating room procedure is performed.
IM03	Cholelitiasis	574.00-574.91	No operating room procedure is performed.
IM04	Heart failure	428.0; 428.1; 428.2; 428.3; 428.4; 428.9	No operating room procedure is performed. Excludes: hypertensive heart failure; rheumatic heart failure.
IM05	Inguinal hernia	550.00-550.03; 550.10-550.13; 550.90-550.93	No operating room procedure is performed.
IM06	Malignant neoplasm of breast	174.0-174.9; 175.0; 175.9	No operating room procedure is performed. Excludes: carcinoma in situ of breast.
IM07	Malignant neoplasm of bronchus and lung	162.2; 162.3; 162.4; 162.5; 162.8; 162.9	No operating room procedure is performed. Excludes: carcinoma in situ of bronchus and lung.
IM08	Normal delivery	650	No operating room procedure is performed. Delivery requiring minimal or no assistance, with or without episiotomy, without fetal manipulation [e.g., rotation version] or instrumentation [forceps] of a spontaneous, cephalic, vaginal, full-term, single, live-born infant[43]
IM09	Pneumonia	480.0; 480.1; 480.2; 480.3; 480.8; 480.9; 481; 482.0; 482.1; 482.2; 482.3; 482.4; 482.8; 482.9; 483.0; 483.1; 483.8; 484.1; 484.3; 484.5; 484.6; 484.7; 484.8; 485; 486	No operating room procedure is performed. Excludes: rheumatic pneumonia.

ICD-9-CM codes and rules by inpatient surgical (IS) case types

[43] Definition reported in the ICD-9-CM FY 2008 code book.

Case type number	*Case type description*	*ICD-9-CM codes*	*Rules*
IS01	Aortic resection, replacement or anastomosis	38.34; 38.44; 39.71; 39.73	Any principal diagnosis code.
IS02	Appendectomy	47.01; 47.09; 47.11; 47.19	Any principal diagnosis code. Includes incidental appendectomy
IS03	Caesarean section	74.0; 74.1; 74.2; 74.4; 74.99	Any principal diagnosis code.
IS04	Cholecystectomy	51.21; 51.22; 51.23; 51.24	Any principal diagnosis code. Includes partial cholecystectomy
IS05	Colorectal resection	45.71; 45.72; 45.73; 45.74; 45.75; 45.76; 45.79; 45.8; 48.41; 48.49; 48.5; 48.61; 48.62; 48.63; 48.64; 48.65; 48.66; 48.69	Any principal diagnosis code.
IS06	Coronary artery bypass graft	36.10; 36.11; 36.12; 36.13; 36.14; 36.15; 36.16; 36.17; 36.19	Any principal diagnosis code.
IS07	Defibrillator insertion, revision, replacement, and removal	00.51; 00.52; 00.54; 37.94; 37.95; 37.96; 37.97; 37.98	Any principal diagnosis code.
IS08	Discectomy	80.50; 80.51; 80.59	Any principal diagnosis code.
IS09	Endarterectomy: vessels of head and neck	38.11; 38.12	Any principal diagnosis code.
IS10	Evacuation of subdural haematoma and intracranial haemorrhage	01.31; 01.39	Any principal diagnosis code.
IS11	Exstirpation, excision and destruction of intracranial lesion	01.41; 01.42; 01.51; 01.52; 01.53; 01.59	Any principal diagnosis code.
IS12	Hip replacement: total and partial	00.70; 00.71; 00.72; 00.73; 81.51; 81.52; 81.53	Any principal diagnosis code. Includes revision of hip replacement.
IS13	Hysterectomy: abdominal and vaginal	68.31; 68.39; 68.41; 68.49; 68.51; 68.59; 68.61; 68.69; 68.71; 68.79; 68.9	Any principal diagnosis code.
IS14	Knee replacement	00.80; 00.81; 00.82; 00.83; 00.84; 81.54; 81.55	Any principal diagnosis code. Includes revision of knee replacement.
IS15	Lumpectomy, quadrantectomy of breast	85.20; 85.21; 85.22; 85.23	Any principal diagnosis code.

Case type number	*Case type description*	*ICD-9-CM codes*	*Rules*
IS16	Mastectomy	85.41; 85.42; 85.43; 85.44; 85.45; 85.46; 85.47; 85.48	Any principal diagnosis code.
IS17	Open prostatectomy	60.3; 60.4; 60.5; 60.61; 60.62; 60.69	Any principal diagnosis code.
IS18	Pacemaker insertion, revision, replacement, and removal	00.50; 00.52; 00.53; 37.70; 37.71; 37.72; 37.73; 37.74; 37.75; 37.76; 37.77; 37.80; 37.81; 37.82; 37.83; 37.85; 37.86; 37.87; 37.89	Any principal diagnosis code. Excludes insertion of temporary pacemaker
IS19	Percutaneous transluminal coronary angioplasty (PTCA)	00.66; 36.01; 36.02; 36.05	Any principal diagnosis code.
IS20	Pheripheral vascular bypass	39.25; 39.29	Any principal diagnosis code.
IS21	Pulmectomy	32.3; 32.4; 32.5	Any principal diagnosis code.
IS22	Repair of inguinal hernia	53.00; 53.01; 53.02; 53.03; 53.04; 53.05; 53.10; 53.11; 53.12; 53.13; 53.14; 53.15; 53.16; 53.17	Any principal diagnosis code.
IS23	Thyroidectomy	06.2; 06.3; 06.4; 06.5; 06.6	Any principal diagnosis code.
IS24	Transurethral resection of prostate (TURP)	60.21; 60.29; 60.96; 60.97	Any principal diagnosis code.

ICD-9-CM codes and rules by outpatient surgical (OS) case types

Case type number	*Case type description*	*ICD-9-CM codes*	*Rules*
OS01	Arthroscopic excision of meniscus of knee	80.26 + 80.6	Any principal diagnosis code. The two codes should be reported at the same time for the same case.
OS02	Breast biopsy and other diagnostic procedures on breast	85.11; 85.12; 85.19	Any principal diagnosis code.
OS03	Cataract surgery	13.11; 13.19; 13.2; 13.3; 13.41; 14.42; 13.43; 13.51; 13.59; 13.64; 13.65; 13.66; 13.69; 13.70; 13.71; 13.72; 13.8; 13.9	Any principal diagnosis code.
OS04	Colonoscopy and biopsy	45.23; 45.24; 45.25	Any principal diagnosis code.
OS05	Colposcopy	70.21	Any principal diagnosis code.

Case type number	*Case type description*	*ICD-9-CM codes*	*Rules*
OS06	Diagnostic bronchoscopy and biopsy of bronchus	33.22; 33.23; 33.24; 33.26; 33.27	Any principal diagnosis code.
OS07	Hysteroscopy	68.12; 68.16	Any principal diagnosis code.
OS08	Ligation and stripping of varicose veins – lower limb	38.59	Any principal diagnosis code.
OS09	Needle biopsy of prostate	60.11	Any principal diagnosis code.
OS10	Proctoscopy and anorectal biopsy	48.23; 48.24; 48.26; 49.21; 49.22; 49.23	Any principal diagnosis code.
OS11	Tonsillectomy and/or adenoidectomy	28.2; 28.3; 28.4; 28.6; 28.7	Any principal diagnosis code.

REFERENCES

Abraham, K.G. and Mackie, C. (eds.) (2005) *Beyond the Market: Designing Nonmarket accounts for the United States*, Committee on National Statistics Division of Behavioural and Social Sciences Education, National Academies Press, Washington DC.

Agence Technique de l'Information sur l'Hospitalisation (2007), Echelle Nationale de couts 2003-2004. (available at: www.atih.sante.fr).

Aizcorbe, A.M., Retus, B.A. and Smith, S. (2008) *Toward a Health Care Satellite Account*, BEA Briefing, May.

Atkinson Review (2005) *Measurement of Government Output and Productivity for the National Accounts.* Final Report, Macmillan, P.

Australian Government Department of Health and Ageing (2006) "National Hospital Cost Data Collection," *Hospital Reference Manual* Round 9, *www.health.gov.au*.

Berndt, E.R., Cutler, D.M., Frank, R.G., Griliches, Z., Newhouse, J.P. and Triplett, J.E. (2000) "Medical care price and output" in Culyer, A.J. and Newhouse, J.P. (Eds.) *Handbook in Health Economics* vol.1a., Elsevier.

Berndt, E.R., Cutler, D.M., Frank, R.G., Griliches, Z., Newhouse, J.P. and Triplett, J.E. (2001) "Price Indexes for Medical Goods and Services: An Overview of Measurement Issues" in Cutler, D.M. and Berndt, E.R. (eds.).

Black, S. (1998); "Measuring the Value of Better Schools"; *FRBNY Economic Policy Review*, Vol.4, No. 1, March.

Brathaug, A.L. (2006) "Measurement of Health Output – experiences from the Norwegian National Accounts" Paper prepared for the joint OECD/ONS/Government of Norway Workshop *Measurement of non-market output in education and health* London, 3-5th October.

Brazier, J., Deverill, M., Green, C., Harper, R., Booth, A. (1999) "A review of the use of health status measures in economic evaluation" *Health Technology Assessment* 3(9). (Available: http://www.hta.nhsweb.nhs.uk/execsumm/summ309.htm).

Busse, R., Schreyogg, J. and Smith, P.C. (2008), "Variability in healthcare treatment costs amongst nine EU countries – results from the health basket project", *Health Economics*; 17(suppl. 1) S1-S8.

Canadian Institute of Health Information (2007), "Day Procedure Groups", http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=casemix_dpg_e.

Cardenas, E.M. (1996), "Revision of the CPI hospital service component", *Monthly Labor Review*, December 1996, pp. 40 – 48.

Castelli, A., Dawson, D., Gravelle, H. And Street, A. (2007) "Improving the Measurement of Health System Output Growth" *Health Economics*16(10), 1091-107.

Castelli, A., Mauro, L. and Street, A. (2008). "Measuring NHS Output Growth", CHE Research Paper 43, October. Available at: http://www.york.ac.uk/inst/che/publications/publicationsbyyear.htm.

Chandler, M. (2009); "An Economic Approach to the Measurement of Growth in the Output of Public Services"; *Economic and Labour Market Review*, Vol. 3, No 10.

Chandler, I.R., Fetter, R.B. and Newbold, R.C. (1991), "Cost Accounting and Budgeting", in Fetter, R.B. (ed.), *DRGs. Their Design and Development*, Health Administration Press, A. Arbour.

Chessa, A.G. and Kleima, F.J. (2006) "The Dutch Experience in Measuring Health Output and Labour Productivity" paper prepared for the joint OECD/ONS/Government of Norway workshop "*Measurement of non-market output in education and health*" London, Brunei Gallery, October 3 – 5, 2006.

Christian, M.S. (2007) "Measuring the Output of Health Care in the United States" *Survey of Current Business*, BEA June, 78-83.

Christian, M.S. (2006) "Measuring the Education Function of Government in the United States"; paper presented at the *OECD/ONS/Government of Norway workshop on Measurement of non-market Output in Education and Health*, London 2006.

Collesi, D, Guerrucci, D., Versace D. and Zannoni, S. (2007); "The use of class size and the Italian method"; paper presented at the *OECD Workshop On Measuring Education and Health Volume Output*, Paris 2007: http://www.oecd.org/document/47/0,3343,en_2825_495684_37733615_1_1_1_1,00.html.

Cutler, D.M., and Berndt, E.R. (editors) (2001); *Medical Care Output and Productivity*; Vol. 62, NBER Studies in Income and Wealth, University of Chicago Press.

Cutler, D.M., McClellan, M., Newhouse, J.P. and Remler, D. (1998); "Are Medical Prices Declining? Evidence from Heart Attack Treatment"; *Quarterly Journal of Economics*, Vol. CXII, November, pp. 991-1023.

Dawson, D. Gravelle, H. et al. (2005) *Developing new approaches to measuring NHS outputs and productivity Final Report;* Centre for Health Economics, University of York, Research Paper 6.

De Han, J. (2009); "On the Measurement of Secondary Education Output in the National Accounts"; paper presented at the UKCeMGA-NIESR conference, November, available under http://www.ons.gov.uk/about/newsroom/events/ukcemga-and-niesr-conference/ukcemga-and-niesr-international-conference-papers/index.html

Department of Health, United Kingdom. (2007), *NHS Costing Manual 2006 – 2007,* www.dh.gov.uk.

Department of Health UK (2008), "Payment by results guidance 2007/2008", *www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_063684*.

Department of Human Services Victoria Australia (n.d.), "Victorian Ambulatory Classification and Funding System – VACS", *www.dhs.vic.gov.au/ahs/archive/vacs/1.htm,* accessed 1 February 2008.

Deveci, N., Heurlén, K. and Sorensen, H.S. (2008); "Non-Market Health Care Service in Denmark – Empirical Studies of A, B and C Methods"; paper presented at the 30th meeting of the *International Association for Research in Income and Wealth*; Portoroz, Slovenia; available under http://www.iariw.org.

Diewert, E.W. (2008); "The Measurement of Nonmarket Sector Outputs and Inputs Using Cost Weights", *University of British Columbia, Department of Economics Discussion Papers* 08-03; available under http://www.econ.ubc.ca/diewert/dp0803.pdf.

Dolan, P. (2000) "The Measurement of Health-related Quality of Life" in Culyer, A.J. and Newhouse, J.P. (eds.) *Handbook of Health Economics* Vol. 1b, ch. 32, Elsevier: Amsterdam.

Ellis, R.P., and Ash, A. (1995) Refinements to the Diagnostic Cost Group (DCG) Model. *Inquiry,* 32 (4): 418-429.

European Hospital and Healthcare Federation. (2006), *DRGs as a Financing Tool*, Brussels, www.hope.be.

Eurostat (1995); *European System of Accounts*; European Commission.

Eurostat (2001); *Handbook on Price and Volume Measures in National Accounts*; European Commission.

Eurostat-OECD (2006), *Methodological Manual on Purchasing Power Parities*, OECD, Paris.

Fernström, M. (2002) "Development of a DRG system for outpatient procedures (NordDRG-O)", *Casemix* vol 4, no 1, March 31.

Fetter, R.B., Thomson, J.D. and Mills, R.E. (1976) "A system for cost and reimbursement control in hospitals", *Yale Journal of Biology and Medicine*, Vol. 49, pp. 123 – 136.

Fisher, F.M. and Shell, K. (1972); *The Economic Theory of Price Indices: Two Essays on the Effects of Taste, Quality, and Technological Change*; New York, Academic Press.

Fries, B.E, Schneider, D.P., Foley, W.J., et al (1994) "Refining a case-mix measure for Nursing Homes: Resource Utilisation Groups (RUG-III)" *Medical Care* 32 (7): 668-685.

Gerard, K, and Mooney, G. (1993) "QALY League Tables: Handle with Care" *Health Economics*, 2: 59-64.

Häkkinen, U. and Joumard, I. (2007) "Cross Country Analysis in OECD Health Care Sectors: Options for Research", *Economics Department Working Paper* no. 554. OECD.

Heikkinen, J. and Hautakangas, S. (2007); "The use of credits and fields of education in tertiary education - the Finnish method"; paper presented at the *OECD Workshop on Measuring Education and Health Volume Output* - 6/7 June.

Heijink, R., Koopmanschap, M.A. and Polder, J.J. (2006) *International Comparisons of Cost of Illness* RIVM, 2006.

Hill, P. (1975); *Price and Volume Measures for Non-Market Services*; Statistical Office of the European Communities, Brussels; available under http://siteresources.worldbank.org/ICPINT/Resources/Price_and_Volume_Measures.pdf.

Hill, P. (1979); *Manual on National Accounts at Constant Prices*; Statistical Papers of the United Nations, Series M, No. 64.

Huber, M. (2007), "International comparison of prices and volumes in health care among OECD countries, European Center for Social Welfare Policy and Research", paper presented at the first meeting of the Task Force for the Development of Health-specific Purchasing Power Parities, Paris, 8 June.

Hughes, J.S., Averill, R.F., Eisenhandler, J., Goldfield, N.I., Muldoon, J., Neff, J.M., and Gay, J.C. (2004) Clinical Risk Groups (CRGs): A Classification System for Risk-Adjusted Capitation-Based Payments and Health Care Management. *Medical Care*, 42 (1): 81-90.

ILO, IMF, OECD, Eurostat, United Nations, World Bank (2004a); *Producer Price Index Manual, Theory and Practice*, Washington, D.C.

ILO, IMF, OECD, Eurostat, United Nations, World Bank (2004b); *Consumer Price Index Manual, Theory and Practice*, Washington, D.C.

Jacobs, R., Smith, P.C. and Street, A. (2006) *Measuring Efficiency in Health Care*, Cambridge University Press.

Jorgenson, D.W. and Fraumeni, B. (1992); "The Output of the Eduation Sector" in Zvi Griliches (ed.) *Output Measurement in the Service Sectors*, Chicago, National Bureau of Economic Research, Studies in Income and Wealth, Vol. 56, University of Chicago Press.

Kronick, R., Gilmer, T., Dreyfus, T., and Lee, L. (2000) Improving Health-Based Payment for Medicaid beneficiaries: CDPS. *Health Care Financing Review*, 21 (3): 29-64.

Krueger, A. (2003); "Economic considerations and class size", *Economic Journal*, 113, pp. F34-F63.

Lee, P. and Little, C.E. (2006) "Measuring health output and productivity in the UK: an essential element of public accountability" Paper prepared for the joint OECD/ONS/Government of Norway Workshop "Measurement of non-market output in education and health" London, Brunei Gallery, Oct 3-5, 2006.

Lohr, K.N. (Ed.) (1990) *Medicare: A Strategy for Quality Assurance* Washington D.C.: National Academy Press.

Magee, H.F. (2003), "The Hospital Data Project. Comparing hospital activity within Europe", *European Journal of Public Health*; Vol. 13 (3 supplement), pp. 73 – 79.

Majeed, A., Bindman, A.B., and Weiner, J.P. (2001) Use of risk adjustment in setting budgets and measuring performance in primary health care I: how it works. *BMJ*, 323: 604-607.

Mattke, S., Kelley, E. et al. (2006) Health Care Quality Indicators Project Initial indicators Report OECD *Health Working Papers* no. 22.

Medpac (2006) *Outpatient Hospital Services Payment System*, September, available at http://www.medpac.gov.

Medpac (2007) *Skilled Nursing Facility services payment system*, October, available at http://www.medpac.gov.

Ministère de la sante et des solidarités (2006), *Manuel des groups homogènes de malades*, France, Paris.

Mogyorosy, Z. and Smith, P. (2005) *The main methodological issues in costing health care services*, Centre for Health Economics, University of York.

Murray, R. (2007); "Developing a quality adjusted output measure for the Scottish Education System"; paper presented at the *OECD Workshop on Measuring Education and Health Volume Output* - 6/7 June.

Nordhaus, W.D. (2002)., "The Health of Nations: The Contribution of Improved Health to Living Standards" (February 2002). *Cowles Foundation Discussion Paper No. 1355*. Available at SSRN: http://ssrn.com/abstract=302789.

Office of Health Economics (2008) NHS Outcomes, Performance and Productivity, March, available at http://www.ohe.org.

OECD (2000), *A System of Health Accounts*, Paris.

OECD (2007) "A selected list of hospital procedures for international comparisons: Report on work of the Expert Group on procedures under HDP2 project" DELSA/HEA/HD/2007(8).

OECD (2009); *Measuring Capital: Revised Manual*; Paris.

OECD (2010); *Handbook on Deriving Capital Measures of Intellectual Property Products*; Paris.

O'Mahoney, M. and Steven, Ph. (2003); *International Comparisons of Performance in the Provision of Public Services: Outcome Based Measures for Education*; National Institute of Economic and Social Research.

ONS (2004) *Public Service Productivity: Health*, Office for National Statistics, October, available at http://www.statistics.gov.uk/CCI/article.asp?ID=1030&Pos=2&ColRank=1&Rank=224.

ONS (2006) *Public Service Productivity: Health*, Office for National Statistics, February, available at http://www.statistics.gov.uk/CCI/article.asp?ID=1030&Pos=2&ColRank=1&Rank=224.

ONS (2008) *Public Service Productivity: Health Care.* Available at: http://www.statistics.gov.uk/cci/article.asp?id=1922.

ONS (2009); Health Care Output 1995-2007; June, available at: http://www.statistics.gov.uk/cci/article.asp?id=2213.

Oostenbrink, J.B. and Rutten, F.F.H. (2006) "Cost assessment and price setting of inpatient care in the Netherlands. The DBC case-mix system" *Health Care Management Science* 9: 287-294.

Roger France, F.H. (2003), "Case mix use in 25 countries: a migration success but international comparisons failure", *International Journal of Medical Informatics*, Vol. 70, pp. 215 – 219.

Rosen, S. (1974); "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition"; *Journal of Political Economy*; 82(1), pp. 34-55.

Rosen, A.K., Reid, R., Broemeling, A., and Rakowski, C.C. (2003) Applying a Risk-Adjustment Framework to Primary Care: Can We Improve on Existing Measures? *Annals of Family Medicine*, 1(1): 44-51.

Schreyer, P. (2008); "Output and Outcome – Measuring the Production of Non-market Services"; paper presented at the 30th meeting of the *International Association for Research in Income and Wealth*; Portoroz, Slovenia; available under http://www.iariw.org.

Schreyogg, J., Tiemann, O., Stargardt, T. and Busse, R. (2008); "Cross-Country Comparisons of Costs: The Use of Episode-Specific Transitive Purchasing Power Parities with Standardised Cost Categories"; *Health Economics*, Vol. 17, pp. 95-103.

Schreyogg, J., Stargardt, T. et al.(2006) "Methods to determine reimbursement rates of diagnosis related groups (DRG): a comparison of nine European countries". Health Care Management Science 9: 215-223.

Schreyogg, J., *et al*. (2005), "Defining the 'Health Benefit Basket' in nine European countries", *European Journal of Health Economics*, Vol. 6 (suppl 1), S2-S10.

Schuster, M.A., McGlynn, E.A. and Brook, R.H. (2005) "How Good is the Quality of Care in the United States" *The Milbank Quarterly*, 83: 843-95.

Scitovski, A.A. (1967), "Changes in the Cost of Treatment of Selected Illnesses, 1951-65"; *American Economic Review*, No. 57, pp. 182-1195.

Smedby, B. (2007), "A selected list of hospital procedures for international comparison. Report on the work of the expert group on procedures under the HDP2 project", presentation at the OECD Health Data National Correspondents meeting, Paris, 9-10 October.

Smith, P.C. and Street, A. (2007) "Measurement of Non-Market Output in Education and Health" Centre for Health Economics, University of York, *CHE Research Paper* 23.

Statistisches Bundesamt (2008); "Preis- und Volumensmessung für Krankenhäuser, Erziehung und Unterricht sowie für öffentliche Sicherheit und Ordnung"; Weiterentwicklung der Methoden, Wiesbaden; available under https://www-ec.destatis.de/csp/shop/sfg/bpm.html.cms.cBroker.cls?cmspath=struktur,vollanzeige.csp&ID=1022668.

Statistics Austria (2006) "ÖNACE 85 – Health and social work".

The Canadian Institute of Health Information "Day Procedure Groups": http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=casemix_dpg_e.

Triplett, J.E. (2009); "Health Care Productivity"; in *Oxford Handbook of Health Economics;* forthcoming.

Triplett, J.E. (2006); Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products; OECD Paris.

Triplett, J.E. (2003), "Integrating Cost-of-disease Studies into Purchasing Power Parities", *A Disease-Based Comparison of Health Systems*, OECD Paris.

Triplett, J.E. (2001); "What's Different About Health? Human Repair and Car Repair in National Accounts and in National Health Accounts"; in Cutler, D. and Berndt, E. (eds.).

Triplett, J.E. (1999) "Accounting for Health Care: Integrating Price index and Cost-Effectiveness Research", in: Triplett, J.E. (ed.); *Measuring the Prices of Medical Treatments*; Brookings Institution.

United Nations et al. 1993, *Systems of National Accounts.*

Weiner, J.P., Starfield, B.H., Steinwachs, D.M., and Mumford, L.M. (1991) Development and Application of a Population-Oriented Measure of Ambulatory Care Case-Mix. Medical Care, 29 (5): 452-472.

Wiley, M.M. (1993), "Costing Hospital Case-Mix: The European Experience", in Casas, M. and Wiley, M.M. (eds.), *Diagnosis Related Groups in Europe*, Springer-Verlag, Berlin.

Williams A. (1985) "Economics of coronary artery bypass grafting" *British Medical Journal* 291: 326-9.

Willmer R, and Little C (2007), Review of data sources and methodology for the calculation of hospital output in the NHS, December, available at http://www.dh.gov.uk/publications.

World Health Organization (2006), *International Shortlist for Hospital Morbidity Tabulation,* *http://who.int/classifications/apps/icd/implementation/hospitaldischarge.htm*, accessed 24 November 2006.