Performance Measurement for Health System Improvement

Experiences, Challenges and Prospects

Peter C. Smith, Elias Mossialos, Irene Papanicolas and Sheila Leatherman

HEALTH ECONOMICS, POLICY AND MANAGEMENT

European Observatory on Health Systems and Policies

WORLD HEALTH ORGANIZATION EUROPE

CAMBRIDGE

PART II

Dimensions of performance

2.1 *Population health*

ELLEN NOLTE, CHRIS BAIN, MARTIN MCKEE

Introduction

Health systems have three goals: (i) to improve the health of the populations they serve; (ii) to respond to the reasonable expectations of those populations; and (iii) to collect the funds to do so in a way that is fair (WHO 2000). The first of these has traditionally been captured using broad measures of mortality such as total mortality, life expectancy, premature mortality or years of life lost. More recently these have been supplemented by measures of the time lived in poor health, exemplified by the use of disability-adjusted life years (DALYs).

These measures are being employed increasingly as a means of assessing health system performance in comparisons between and within countries. Their main advantage is that the data are generally available. The most important drawback is the inability to distinguish between the component of the overall burden of disease that is attributable to health systems and that which is attributable to actions initiated elsewhere. *The world health report 2000* sought to overcome this problem by adopting a very broad definition of a health system as "all the activities whose primary purpose is to promote, restore or maintain health" (WHO 2000) (Box 2.1.1). A somewhat circular logic makes it possible to use this to justify the use of DALYs as a measure of performance. However, in many cases policy-makers will wish to examine a rather more narrow question – how is a particular health system performing in the delivery of health care?

This chapter examines some of these issues in more detail. It does not review population health measurement per se, as this has been addressed in detail elsewhere (see, for example, Etches et al. 2006; McDowell et al. 2004; Murray et al. 2000; Murray et al. 2002; Reidpath 2005). However, we give a brief overview of some measures that have commonly been used to assess population health in relation

Box 2.1.1 Defining health systems

Many activities that contribute directly or indirectly to the provision of health care may or may not be within what is considered to be the health system in different countries (Nolte et al. 2005). Arah and colleagues (2006) distinguish between the *health* system and the *health-care* system. The latter refers to the "combined functioning of public health and personal health-care services" that are under the "direct control of identifiable agents, especially ministries of health." In contrast, the health system extends beyond these boundaries "to include all activities and structures that impact or determine health in its broadest sense within a given society". This closely resembles the World Health Organization (WHO) definition of a health system set out in *The world health report 2000* (WHO 2000). Consequently, health-care performance refers to the "maintenance of an efficient and equitable system of health care", evaluating the system of health-care delivery against the "established public goals for the level and distribution of the benefits and costs of personal and public health care" (Arah et al. 2006). Health system performance is based on a broader concept that also takes account of determinants of population health not related to health care, principally building on the health field concept advanced by Lalonde and thus subsuming health-care performance (Lalonde 1974).

to health-care performance (Annex 1 & 2). We begin with a short historical reflection of the impact of health care on population health. We discuss the challenges of attributing population health outcomes to activities in the health system, and thus of identifying indicators of health system performance, before considering indicators and approaches that have been developed to relate measures of health at the population level more closely to health-care performance.

Does health care contribute to population health?

There has been long-standing debate about whether health services make a meaningful contribution to population health (McKee 1999). Writing from a historical perspective in the late 1970s, several authors argued that health care had contributed little to the observed decline in

mortality that had occurred in industrialized countries from the mid-nineteenth to the mid-twentieth century. It was claimed that mortality improvements were most likely to be attributable to the influence of factors outside the health-care sector, particularly nutrition, but also to general improvements in the environment (Cochrane et al. 1978; McKeown 1979; McKinlay & McKinlay 1977).

Much of this discussion has been linked to the work of Thomas McKeown (Alvarez-Dardet & Ruiz 1993). His analysis of the mortality decline in England and Wales between 1848/1854 and 1971 illustrated how the largest part of an observed fall in death rates from tuberculosis (TB) predated the introduction of interventions such as immunization or effective chemotherapy (McKeown 1979). He concluded that "specific measures of preventing or treating disease in the individual made no significant contribution to the reduction of the death rate in the nineteenth century" (McKeown 1971), or indeed into the mid-twentieth century. His conclusions were supported by contemporaneous work which analysed long-term trends in mortality from respiratory TB until the early and mid-twentieth century in Glasgow, Scotland (Pennington 1979); and in England and Wales, Italy and New Zealand (Collins 1982); and from infectious diseases in the United States of America in the early and mid-twentieth century (McKinley & McKinley 1977).

Recent reviews of McKeown's work have challenged his sweeping conclusions. They point to other evidence, such as that which demonstrated that the decline in TB mortality in England and Wales in the late nineteenth and early twentieth centuries could be linked in part to the emerging practice of isolating poor patients with TB in workhouse infirmaries (Fairchild & Oppenheimer 1998; Wilson 2005). Nolte and McKee (2004) showed how the pace at which mortality from TB declined increased markedly following the introduction of chemotherapy in the late 1940s, with striking year-on-year reductions in death rates among young people. Others contended that McKeown's focus on TB may have overstated the effect of changing living standards and nutrition (Szreter 1988) and simultaneously underestimated the role of medicine. For example, the application of inoculation converted smallpox from a major to a minor cause of death between the late eighteenth and early nineteenth centuries (Johansson 2005).

Similarly, Schneyder and colleagues (1981) criticized McKinley and McKinley's (1977) analysis for adopting a narrow interpreta-

tion of medical measures, so disregarding the impact of basic public health measures such as water chlorination. Evidence provided by Mackenbach (1996), who examined a broader range of causes of death in the Netherlands between 1875/1879 and 1970, also suggests that health care had a greater impact than McKeown and others had acknowledged. Mackenbach (1996) correlated infectious disease mortality with the availability of antibiotics from 1946 and deaths from common surgical and perinatal conditions with improvements in surgery and anaesthesia and in antenatal and perinatal care since the 1930s. He estimated that up to 18.5% of the total decline in mortality in the Netherlands between the late nineteenth and mid-twentieth centuries could be attributed to health care.

However, this debate does not address the most important issue. McKeown was describing trends in mortality at a time when health care could, at best, contribute relatively little to overall population health as measured by death rates. Colgrove (2002) noted that there is now consensus that McKeown was correct to the extent that "curative medical measures played little role in mortality decline prior to the mid-20th century." However, the scope of health care was beginning to change remarkably by 1965, the end of the period that McKeown analysed. A series of entirely new classes of drugs (for example, thiazide diuretics, beta blockers, beta-sympathomimetics, calcium antagonists) made it possible to control common disorders such as hypertension and chronic airways diseases. These developments, along with the implementation of new and more effective ways of organizing care and the development of evidence-based care, made it more likely that health care would play a more important role in determining population health.

How much does health care contribute to population health?

Given that health care can indeed contribute to population health – how much of a difference does it actually make? Bunker and colleagues (1994) developed one approach to this question, using published evidence on the effectiveness of specific health service interventions to estimate the potential gain in life expectancy attributable to their introduction. For example, they examined the impact of thirteen clinical preventive services (such as cervical cancer screening) and thirteen curative services (such as treatment of cervical cancer) in the United States and estimated

a gain of eighteen months from preventive services. A potential further gain of seven to eight months could be achieved if known efficacious measures were made more widely available. The gain from curative services was estimated at forty-two to forty-eight months (potential further gain: twelve to eighteen months). Taken together, these calculations suggest that about half of the total gain in life expectancy (seven to seven and a half years) in the United States since 1950 may be attributed to clinical preventive and curative services (Bunker 1995).

Wright and Weinstein (1998) used a similar approach to look at a range of preventive and curative health services but focused on interventions targeted at populations at different levels of risk (average and elevated risk; established disease). For example, they estimated that a reduction in cholesterol (to 200 mg/dL) would result in life expectancy gains of fifty to seventy-six months in thirty-five year-old people with highly elevated blood cholesterol levels (> 300 mg/dL). In comparison, it was estimated that life expectancy would increase by eight to ten months if average-risk smokers aged thirty-five were helped to stop smoking.

Such analyses provide important insights into the potential contribution of health care to population health. However, they rest on the assumption that the health gains reported in clinical trials translate directly to the population level. This is not necessarily the case (Britton et al. 1999) as trial participants are often highly selected subsets of the population, typically excluding elderly people and those with comorbidities. Also, evaluations of individual interventions fail to capture the combined effects of integrated and individualized packages of care (Buck et al. 1999). The findings thus provide little insight into what health systems actually achieve in terms of health gain or how different systems compare.

An alternative approach uses regression analysis to identify any link between inputs to health care and health outcomes although such studies have produced mixed findings. Much of the earlier work failed to identify strong and consistent relationships between health-care indicators (such as health-care expenditure, number of doctors) and health outcomes (such as (infant) mortality, life expectancy) but found socio-economic factors to be powerful determinants of health outcomes (Babazono & Hillman 1994; Cochrane et al. 1978; Kim & Moody 1992). More recent work has provided more consistent evidence. For example, significant inverse relationships have been established between health-care expenditure and infant and premature

mortality (Cremieux et al. 1999; Nixon & Ulmann 2006; Or 2000); and between the number of doctors per capita and premature and infant mortality, as well as life expectancy at age sixty-five (Or 2001).

Other studies have asked whether the organization of health-care systems is important. For example, Elola and colleagues (1995), and van der Zee and Kroneman (2007) studied seventeen health-care systems in western Europe. They distinguished national health service (NHS) systems (such as those in Denmark, Ireland, Italy, Spain, United Kingdom) from social security systems (such as those in Germany, Austria, the Netherlands). Controlling for socio-economic indicators and using a cross-sectional analysis, Elola and colleagues (1995) found that countries with NHS systems achieve lower infant mortality rates than those with social security systems at similar levels of gross domestic product (GDP) and health-care expenditure. In contrast, van der Zee and Kroneman (2007) analysed long-term time trends from 1970 onwards. They suggest that the relative performance of the two types of systems changed over time and social security systems have achieved slightly better outcomes (in terms of total mortality and life expectancy) since 1980, when inter-country differences in infant mortality became negligible.

These types of study have obvious limitations arising from data availability and reliability as well as other less-obvious limitations. One major weakness is the cross-sectional nature that many of them display. Gravelle and Blackhouse (1987) have shown how such analyses fail to take account of lagged relationships. An obvious example is cancer mortality, in which death rates often reflect treatments undertaken up to five years previously. Furthermore, a cross-sectional design is ill-equipped to address adequately causality and such models often lack any theoretical basis that might indicate what causal pathways may exist (Buck et al. 1999). However, the greatest problem is that the majority of studies of this type employ indicators of population health (for example, life expectancy and total mortality) that are influenced by many factors outside the health-care sector. These include policies in sectors such as education, housing and employment, where the production of health is a secondary goal.

This is also true of more restricted measures of mortality. Thus, infant mortality rates are often used in international comparisons to capture health-care performance. Yet, deaths in the first four weeks of life (neonatal) and those in the remainder of the first year (postneo-

natal) have quite different causes. Postneonatal mortality is strongly related to socio-economic factors while neonatal mortality more closely reflects the quality of medical care (Leon et al. 1992). Consequently, assessment of the performance of health care per se requires identification of the indicators of population health that most directly reflect that care.

Attributing indicators of population health to activities in the health system

As noted in the previous section, the work by Bunker and colleagues (1994) points to a potentially substantial contribution of health care to gains in population health, although that contribution has not been quantified. In some cases the impact of health care is almost self-evident, as is the case with vaccine-preventable disease. This is illustrated by the eradication of smallpox in 1980 that followed systematic immunization of entire populations in endemic countries, and also by antibiotic treatment of many common infections. The discovery of insulin transformed type I diabetes from a rapidly fatal childhood illness to one for which optimal care can now provide an almost normal lifespan. In these cases, observed reductions in mortality can be attributed quite clearly to the introduction of new treatments. For example, there was a marked reduction in deaths from testicular cancer in the former East Germany when modern chemotherapeutic agents became available after unification (Becker & Boyle 1997). In other situations the influence is less clear, particularly when the final outcome is only partly attributable to health care. In this chapter we use the examples of ischaemic heart disease, perinatal mortality and cancer survival to illustrate some of the challenges involved in using single indicators of population health to measure health system performance.

Ischaemic heart disease

Ischaemic heart disease is one of the most important causes of premature death in industrialized countries. Countries in western Europe have had great success in controlling this disease and death rates have fallen, on average, by about 50% over the past three decades (Kesteloot et al. 2006) (Fig. 2.1.1). Many new treatments have been introduced including new drugs for heart failure and cardiac arrhythmias; new

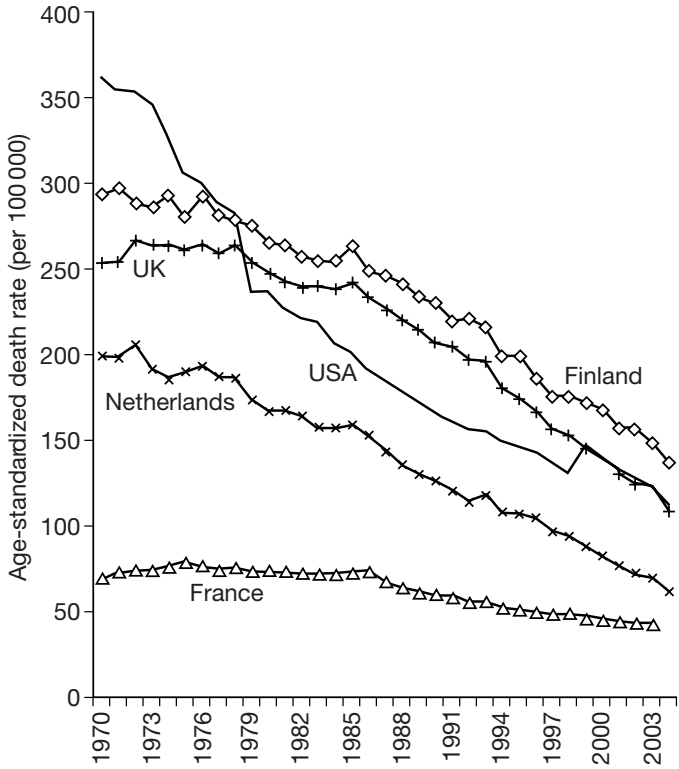

Fig. 2.1.1 Mortality from ischaemic heart disease in five countries, 1970–2004
Source: OECD 2007

technology, such as more advanced pacemakers; and new surgical techniques, such as angioplasty. Although still somewhat controversial, accumulating evidence suggests that these developments have made a considerable contribution to the observed decline in ischaemic heart disease mortality in many countries.

Beaglehole (1986) calculated that 40% of the decline in deaths from ischaemic heart disease in Auckland, New Zealand between 1974 and 1981 could be attributed to advances in medical care. Similarly, a study in the Netherlands estimated that specific medical interventions (treatment in coronary care units, post-infarction treatment, coronary artery bypass grafting (CABG)) had potentially contributed to 46% of the observed decline in mortality from ischaemic heart disease between 1978 and 1985. Another 44% was attributed to primary prevention

efforts such as smoking cessation, strategies to reduce cholesterol levels and treatment of hypertension (Bots & Grobee 1996).

Hunink and colleagues (1997) estimated that about 25% of the decline in ischaemic heart disease mortality in the United States between 1980 and 1990 could be explained by primary prevention and another 72% was due to secondary reduction in risk factors or improvements in treatment. Capewell and colleagues (1999, 2000) assessed the contribution of primary (such as treatment of hypertension) and secondary (e.g. treatment following myocardial infarction) prevention measures to observed declines in ischaemic heart disease mortality in a range of countries during the 1980s and 1990s. Using the IMPACT model, they attributed between 23% (Finland) and almost 50% (United States) of the decline to improved treatment. The remainder was largely attributed to risk factor reductions (Table 2.1.1) (Ford et al. 2007). These estimates gain further support from the WHO Multinational Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA) project which linked changes in coronary care and secondary prevention practices to the decline in adverse coronary outcomes between the mid-1980s and the mid-1990s (Tunstall-Pedoe et al. 2000).

In summary, these findings indicate that between 40% and 50% of the decline in ischaemic heart disease in industrialized countries can be attributed to improvements in health care. Yet, it is equally clear that large international differences in mortality predated the advent of effective health care, reflecting factors such as diet, rates of smoking and physical activity. Therefore, cross-national comparisons of ischaemic heart disease mortality have to be interpreted in the light of wider policies that determine the levels of the main cardiovascular risk factors in a given population (Box 2.1.2).

The nature of observed trends may have very different explanations. This is illustrated by the former East Germany and Poland, which both experienced substantial declines in ischaemic heart disease mortality during the 1990s – reductions of approximately one fifth between 1991/1992 and 1996/1997 among those aged under seventy-five years (Nolte et al. 2002).

In Poland, this improvement has been largely attributed to changes in dietary patterns, with increasing intake of fresh fruit and vegetables and reduced consumption of animal fat (Zatonski et al. 1998). The contribution of medical care was considered to be negligible,

Table 2.1.1 ***Decline in ischaemic heart disease mortality attributable to treatment and to risk factor reductions in selected study populations (%)***

Country	Period	Risk factors	Treatment
Auckland, New Zealand (Beaglehole 1986)	1974–1981	–	40%
Netherlands (Bots & Grobee 1996)	1978–1985	44%	46%
United States (Hunink et al. 1997)	1980–1990	50%	43%
Scotland (Capewell et al. 1999)	1975–1994	55%	35%
Finland (Laatikainen et al. 2005)	1982–1997	53%	23%
Auckland, New Zealand (Capewell et al. 2000)	1982–1993	54%	46%
United States (Ford et al. 2007)	1980–2000	44%	47%
Ireland (Bennett et al. 2006)	1985–2000	48%	44%
England & Wales (Unal et al. 2007)	1981–2000	58%	42%

although data from the WHO MONICA project in Poland suggest that there was a considerable increase in intensity of the treatment of acute coron-ary events between 1986/1989 and the early 1990s (Tunstall-Pedoe et al. 2000). However, Poland has a much higher proportion of sudden deaths from ischaemic heart disease in comparison with the west. This phenomenon has also been noted in the neighbouring Baltic republics and in the Russian Federation (Tunstall-Pedoe et al. 1999; Uuskula et al. 1998) and has been related to binge drinking (McKee et al. 2001). From this it would appear that health care has been of minor importance in the overall decline in ischaemic heart disease mortality in Poland in the 1990s.

The eastern part of Germany experienced substantial increases in a variety of indicators of intensified treatment of cardiovascular disease during the 1990s (for example, cardiac surgery increased by 530%

Box 2.1.2 Comparing mortality across countries

International variations in ischaemic heart disease mortality and, by extension, other cause-specific mortality may be attributable (at least in part) to differences in diagnostic patterns, death certification or cause of death coding in each country. This problem is common to all analyses that employ geographical and/or tem-poral analyses of mortality data. However, it must be set against the advantages of mortality statistics – they are routinely available in many countries and, as death is a unique event (in terms of its finality), it is clearly defined (Ruzicka & Lopez 1990). Of course there are some caveats. Mortality data inevitably underestimate the burden of disease attributable to low-fatality conditions (such as mental illness) or many chronic disorders that may rarely be the immediate cause of death but which contribute to deaths from other causes. For example, diabetes contributes to many deaths from ischaemic heart disease or renal failure (Jougla et al. 1992). Other problems arise from the different steps involved in the complex sequence of events that leads to allocation of a code for cause of death (Kelson & Farebrother 1987; Mackenbach et al. 1987). For example, the diagnostic habits and preferences of certifying doctors are likely to vary with the diagnostic techniques available, cultural norms or even professional training. The validity of cause of death statistics may also be affected by the process of assigning the formal International Classification of Diseases (ICD) code to the statements on the death certificate. However, a recent evaluation of cause of death statistics in the European Union (EU) found the quality and comparability of cardiovascular and respiratory death reporting across the region to be sufficiently valid for epidemiological purposes (Jougla et al. 2001). Where there were perceived problems in comparability across countries, the observed differences were not large enough to explain fully the variations in mortality from selected causes of cardiovascular or respiratory death.

Overall, mortality data in the European region are generally considered to be of good quality, although some countries have been experiencing problems in ensuring complete registration of all deaths. Despite some improvements since the 1990s, problems remain with recent figures estimating completeness of mortality

Box 2.1.2 cont'd

data covered by the vital registration systems range from 60% in Albania; 66% to 75% in the Caucasus; and 84% to 89% in Kazakhstan and Kyrgyzstan (Mathers et al. 2005). Also, the vital registration system does not cover the total resident population in several countries, excluding certain geographical areas such as Chechnya in the Russian Federation; the Transnistria region in Moldova; or Kosovo, until recently part of Serbia (WHO Regional Office for Europe 2007).

between 1993 and 1997) (Brenner et al. 2000). However, intensified treatment does not necessarily translate into improved survival rates (Marques-Vidal et al. 1997). There was a (non-significant) increase in the prevalence of myocardial infarction among people from the east of Germany aged twenty-five to sixty-nine years, between 1990/1992 and 1997/1998, which accompanied an observed decline in ischaemic heart disease mortality, suggesting that the latter is likely to be attributable to improved survival (Wiesner et al. 1999).

In summary, a fall in ischaemic heart disease mortality can generally be seen as a good marker of effective health care and usually contributes to around 40% to 50% of observed declines. However, multiple factors influence the prevalence of ischaemic heart disease. As some lie within the control of the health-care sector and others require intersectoral policies, it may not be sufficient to use ischaemic heart disease mortality as a sole indicator of health-care performance. At the same time, ischaemic heart disease may be considered to be an indicator of the performance of national systems as a whole. Continuing high levels point to a failure to implement comprehensive approaches that cover the entire spectrum – from health promotion through primary and secondary prevention to treatment of established disease.

Perinatal mortality

Perinatal mortality (see Annex 2) has frequently been used as an indicator of the quality of health care (Rutstein et al. 1976). However, comparisons between countries and over time are complicated because rates are now based on very small numbers which are "very dependent on precise definitions of terms and variations in local practices

and circumstances of health care and registration systems" (Richardus et al. 1998). For example, advances in obstetric practice and neonatal care have led to improved survival of very preterm infants. These outcomes affect attitudes to the viability of such infants (Fenton et al. 1992) and foster debate about the merits of striving to save very ill newborn babies (who may suffer long-term brain damage) or making the decision to withdraw therapy (De Leeuw et al. 2000). Legislation and guidelines concerning end-of-life decisions vary among countries – some protect human life at all costs; some undertake active interventions to end life, such as in the Netherlands (McHaffie et al. 1999).

A related problem is that registration procedures and practices may vary considerably between countries, reflecting different legal definitions of the vital events. For example, the delay permitted for registration of births and deaths ranges from three to forty-two days within western Europe (Richardus et al. 1998). This is especially problematic for small and preterm births, as deaths that occur during the first day of life are most likely to be under-registered in countries with the longest permitted delays.

Congenital anomalies are an important cause of perinatal mortality. However, improved ability of prenatal ultrasound screening to recognize congenital anomalies has been shown to reduce perinatal mortality as fetuses with such anomalies are aborted rather than surviving to become fetal or infant deaths (Garne 2001; Richardus et al. 1998). This phenomenon may distort international comparisons (van der Pal-de Bruin et al. 2002). Garne and colleagues (2001) demonstrated how a high frequency of congenital mortality (44%) among infant deaths in Ireland reflected limited prenatal screening and legal prohibition of induced abortion. Conversely, routine prenatal screening in France is linked to ready access to induced abortion throughout gestation. Congenital mortality was cited in 23% of infant deaths although the total number of deaths from congenital malformations (aborted plus delivered) was higher in France (Garne et al. 2001). However, recent work in Italy has demonstrated that the relative proportion of congenital anomalies as a cause of infant deaths tends to remain stable within countries (Scioscia et al. 2007). This suggests that perinatal mortality does provide important insights into the performance of (neonatal) care over time.

In summary, international comparisons of perinatal mortality should be interpreted with caution. However, notwithstanding improvements

in antenatal and obstetric care in recent decades, perinatal audit studies that take account of these factors show that improved quality of care could reduce current levels of perinatal mortality by up to 25% (Richardus et al. 1998). Thus, perinatal mortality can serve as a meaningful outcome indicator in international comparisons as long as care is taken to ensure that comparisons are valid. The EuroNatal audit in regions of ten European countries showed that differences in perinatal mortality rates may be explained in part by differences in the quality of antenatal and perinatal care (Richardus et al. 2003).

Cancer survival

Cancer survival statistics have intrinsic appeal as a measure of health system performance – cancer is common; causes a large proportion of total deaths; and is one of the few diseases for which individual survival data are often captured routinely in a readily accessible format. This has led to their widespread use for cross-sectional assessments of differences within population subgroups (Coleman et al. 1999) and over time (Berrino et al. 2007; Berrino et al. 2001). Comparisons within health systems have clear potential for informing policy by providing insight into differences in service quality, for example: timely access, technical competence and the use of standard treatment and follow-up protocols (Jack et al. 2003).

International comparisons of cancer registry data have revealed wide variations in survival among a number of cancers of adults within Europe. The Nordic countries generally show the highest survival rates for most common cancers (Berrino et al. 2007; Berrino et al. 2001) (Fig. 2.1.2) and there are marked differences between Europe and the United States (Gatta et al. 2000).

Prima facie, these differences might suggest differing quality of care, so cancer survival has been proposed as an indicator of international differences in health-care performance (Hussey et al. 2004; Kelley & Hurst 2006). However, recent commentaries highlight the many elements that influence cancer outcomes (Coleman et al. 1999; Gatta et al. 2000). These include the case-mix, that is, the distribution of tumour stages. These will depend on the existence of screening programmes, as with prostate and breast cancer; the socio-demographic composition of the population covered by a registry (not all registries cover the entire population); and time lags (personal and system induced)

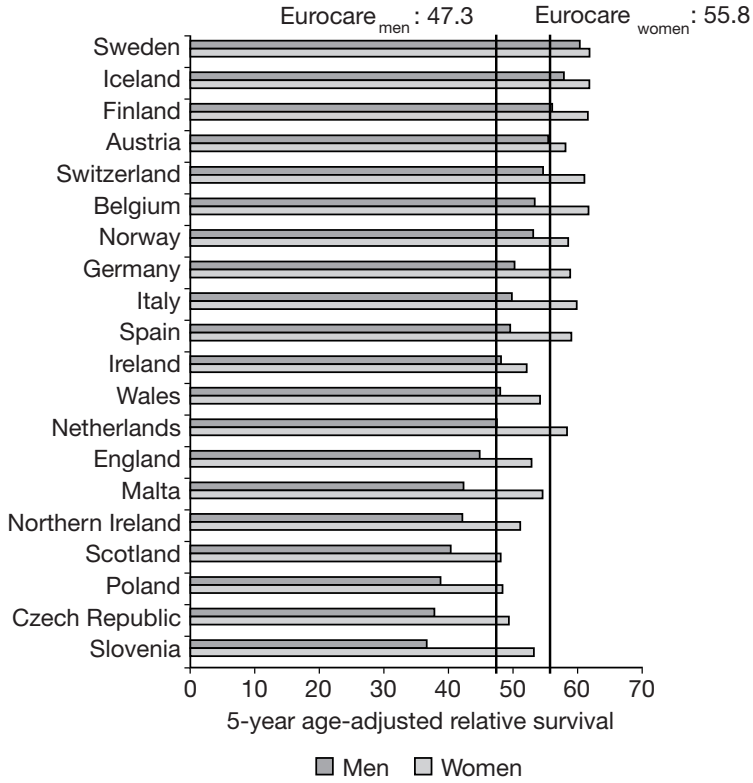

Fig. 2.1.2 Age-adjusted five-year relative survival of all malignancies of men and women diagnosed 2000–2002

Source: Verdecchia et al. 2007

between symptom occurrence and treatment (Sant et al. 2004). Data from the United States suggest that the rather selected nature of the populations covered by the registries of the Surveillance Epidemiology and End Results (SEER) Program, widely used in international comparisons, account for much of the apparently better survival rates in the United States for a number of major cancers (Mariotto et al. 2002). Death rates increased by 15% for prostate cancer; 12% for breast cancer; and 6% for colorectal cancer in men when SEER rates were adjusted to reflect the characteristics of the American population. This brings them quite close to European survival figures.

Presently, routine survival data incorporate adjustments only for age and the underlying general mortality rate of a population.

Use of stage-specific rates would improve comparability (Ciccolallo et al. 2005) but these are not widely available, nor are they effective for comparisons of health systems at different evolutionary stages. A more sophisticated staging system based on intensive diagnostic workup can improve stage-specific survival for all stages – those transferred from the lower stage will usually have lower survival than those remaining in the former group, but better survival than those initially in the higher stage.

Sometimes there is uncertainty about the diagnosis of malignancy (Butler et al. 2005). For example, there is some suggestion that apparently dramatic improvements in survival among American women with ovarian cancer in the late 1980s may be largely attributable to changes in the classification of borderline ovarian tumours (Kricker 2002). The ongoing CONCORD study of cancer survival is examining these issues in detail across four continents, supporting future calibration and interpretation of cancer survival rates (Ciccolallo et al. 2005; Gatta et al. 2000). There is little doubt that survival rates should be considered as no more than a means to flag possible concerns about health system performance at present.

Yet, it is important to note that while cross-national comparisons – whether of cancer survival (illustrated here) or other disease-specific population health outcomes (such as ischaemic heart disease mortality, described earlier) can provide important insights into the relative performance of health-care systems. It will be equally important for systems to benchmark their progress against themselves over time. For example, cross-national comparisons of breast cancer survival in Europe have demonstrated that constituent parts of the United Kingdom have relatively poor performance in comparison with other European countries (Berrino et al. 2007) (Fig. 2.1.3).

However, this has to be set against the very rapid decline in mortality from breast cancer in the United Kingdom since 1990 (Fig. 2.1.4), pointing to the impact of improvements in diagnostics and treatment (Kobayashi 2004). Thus, a detailed assessment of progress of a particular system optimally includes a parallel approach that involves both cross-sectional and longitudinal analyses. In the case of cancer survival these should ideally be stage-specific so as to account for inherent potential biases that occur when short-term survival is used to assess screening effects.

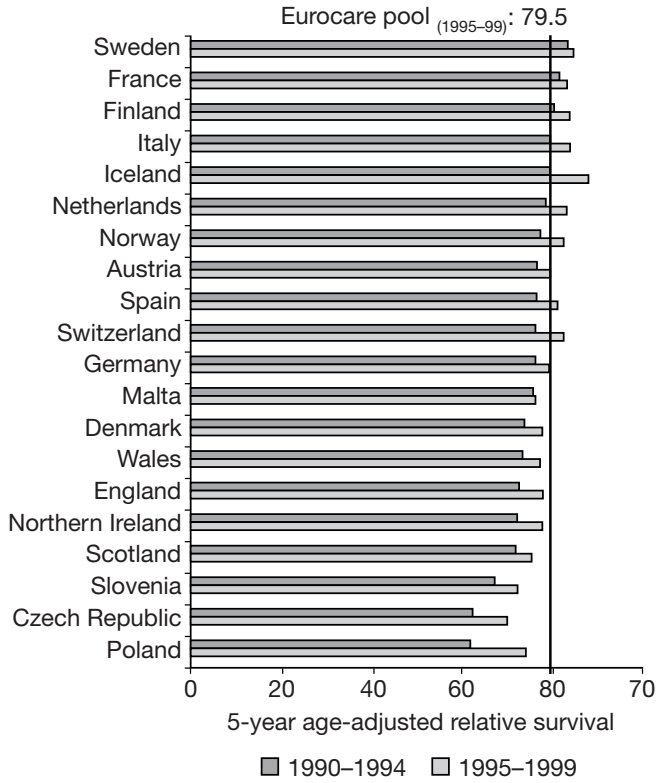

Fig. 2.1.3 Age-adjusted five-year relative survival for breast cancer for women diagnosed 1990–1994 and 1995–1999

Source: Berrino et al. 2007

In summary, these examples of ischaemic heart disease mortality, perinatal mortality and cancer survival indicate the possibilities and the challenges associated with particular conditions. Each provides a lens to examine certain elements of the health-care system. In the next section these are combined with other conditions amenable to timely and effective care to create a composite measure – avoidable mortality.

Concept of avoidable mortality

The concept of avoidable mortality originated with the Working Group on Preventable and Manageable Diseases led by David Rutstein of Harvard Medical School in the United States in the 1970s

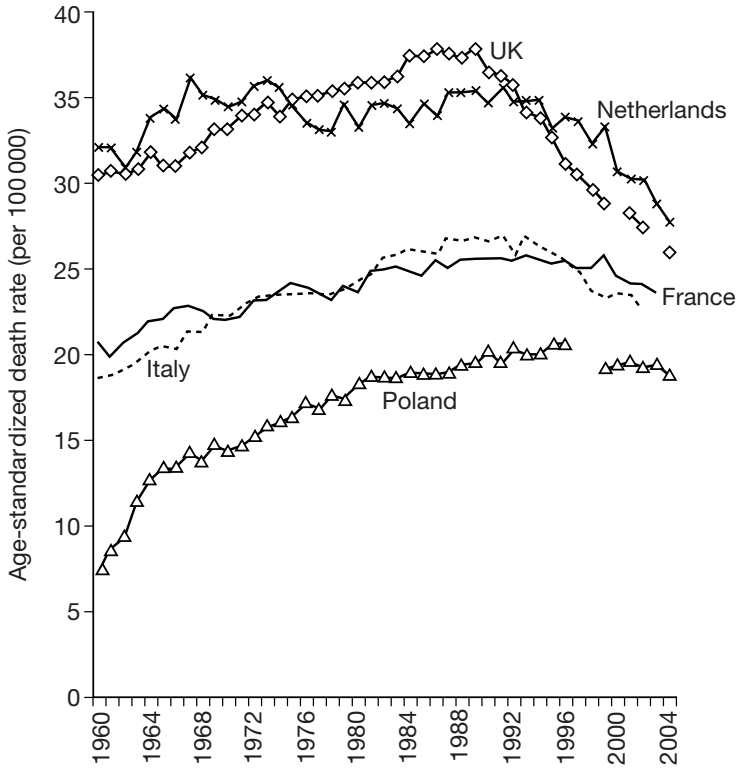

Fig. 2.1.4 Age-standardized death rates from breast cancer in five countries, 1960–2004

Source: OECD 2007

(Rutstein et al. 1976). They introduced the notion of 'unnecessary untimely deaths' by proposing a list of conditions from which death should not occur in the presence of timely and effective medical care. This work has given rise to the development of a variety of terms including 'avoidable mortality' and 'mortality amenable to medical/health care' (Charlton et al. 1983; Holland 1986; Mackenbach et al. 1988). It attracted considerable interest in the 1980s as a way of assessing the quality of health care, with numerous researchers, particularly in Europe, applying it to routinely collected mortality data. It gained momentum with the European Commission Concerted Action Project on Health Services and 'Avoidable Deaths', established in the early 1980s. This led to the publication of the *European Community*

Atlas of Avoidable Death in 1988 (Holland 1988), a major work that has been updated twice.

Nolte and McKee (2004) reviewed the work on avoidable mortality undertaken until 2003 and applied an amended version of the original lists of causes of death considered amenable to health care to countries in the EU (EU15)[1]. They provide clear evidence that improvements in access to effective health care had a measurable impact in many countries during the 1980s and 1990s. Interpreting health care as primary care, hospital care, and primary and secondary preventive services such as screening and immunization, they examined trends in mortality from conditions for which identifiable health-care interventions can be expected to avert mortality below a defined age (usually seventy-five years). Assuming that, although not all deaths from these causes are entirely avoidable, health services could contribute substantially by minimizing mortality but demonstrated how such deaths were still relatively common in many countries in 1980. However, reductions in these deaths contributed substantially to the overall improvement in life expectancy between birth and age seventy-five during the 1980s. In contrast, declines in avoidable mortality made a somewhat smaller contribution to the observed gains in life expectancy during the 1990s, especially in the northern European countries that had experienced the largest gains in the preceding decade.

Importantly, although the rate of decline in these deaths began to slow in many countries in the 1990s, rates continued to fall even in countries that had already achieved low levels. For example, this was demonstrated for 19 industrialized countries between 1997/1998 and 2002/2003, although the scale and pace of change varied (Nolte & McKee 2008) (Fig. 2.1.5). The largest reductions were seen in countries with the highest initial levels (including Portugal, Finland, Ireland, United Kingdom) and also in some countries that had been performing better initially (such as Australia, Italy, France). In contrast, the United States started from a relatively high level of avoidable mortality but experienced much smaller reductions.

The concept of avoidable mortality provides a valuable indicator of general health-care system performance but has several limitations. These have been discussed in detail (Nolte & McKee 2004). We here focus on three aspects that need to be considered when interpreting observed trends: the level of aggregation; the coverage of health

[1] EU15: Member States belonging to the European Union before 1 May 2004.

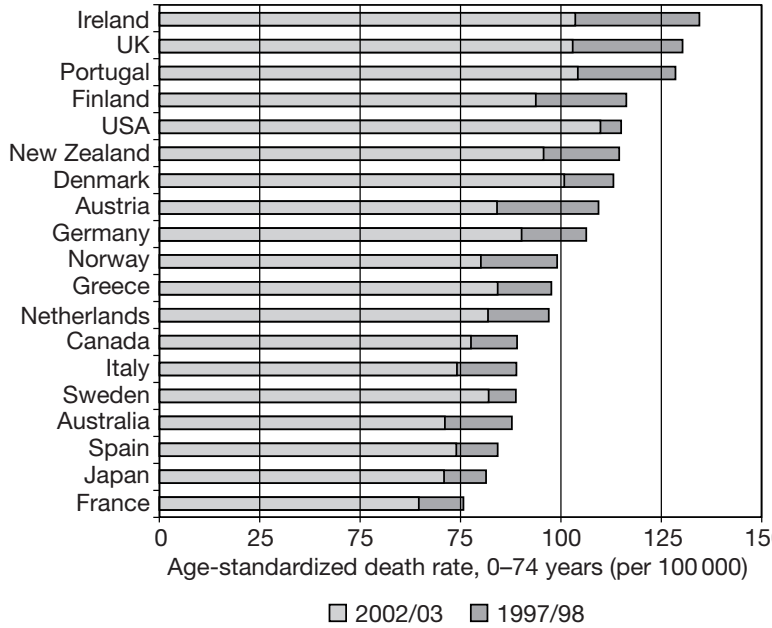

Fig. 2.1.5 Mortality from amenable conditions (men and women combined), age 0–74 years, in 19 OECD countries, 1997/98 and 2002/03 (Denmark: 2000/01; Sweden: 2001/02; Italy, United States: 2002)

Source: Adapted from Nolte & McKee 2008

outcomes; and the attribution of outcomes to activities in the health system.

Nolte and McKee (2008) noted that there are likely to be many underlying reasons for an observed lack of progress on the indicator of amenable mortality in the United States. Any aggregate national figure will inevitably conceal large variations due to geography, race and insurance coverage, among many other factors. Interpretation of the data must go beyond the aggregate figure to look within populations and at specific causes of death if these findings are to inform policy.

The focus on mortality is one obvious limitation of the concept of avoidable mortality. At best mortality is an incomplete measure of health-care performance and is irrelevant for those services that are focused primarily on relieving pain and improving quality of life. However, reliable data on morbidity are still scarce. There has been progress in setting up disease registries other than the more widely

established cancer registries (for example, for conditions such as diabetes, myocardial infarction or stroke) but information may be misleading where registration is not population-based. Population surveys provide another potential source of data on morbidity, although survey data are often not comparable across regions. Initiatives such as the European Health Survey System currently being developed by Eurostat and the European Commission's Directorate-General for Health and Consumers (DG SANCO) will go some way towards developing and collecting consistent indicators (European Commission 2007). Routinely collected health service utilization data such as inpatient data or consultations of general practitioners and/or specialists usually cover an entire region or country. However, while potentially useful, these data (especially consultation rates) do not include those who need care but fail to seek it.

Finally, an important issue relates to the list of causes of death considered amenable to health care. Nolte and McKee (2004) define amenable conditions "[as] those from which it is reasonable to expect death to be averted even after the condition develops". This interpretation would include conditions such as TB, in which the acquisition of disease is largely driven by socio-economic conditions but timely treatment is effective in preventing death. This highlights how the attribution of an outcome to a particular aspect of health care is intrinsically problematic because of the multi-factorial nature of most outcomes. As a consequence, when interpreting findings a degree of judgement, based on an understanding of the natural history and scope for prevention and treatment of the condition in question, is needed. Thus it will be possible to distinguish more clearly between conditions in which death can be averted by health-care intervention (amenable conditions) as opposed to interventions reflecting the relative success of policies outside the direct control of the health-care sector (preventable conditions). Preventable conditions thus include those for which the aetiology is mostly related to lifestyle factors, most importantly the use of tobacco and alcohol (lung cancer and liver cirrhosis). This group also includes deaths amenable to legal measures such as traffic safety (speed limits, use of seat belts and motorcycle helmets). This refined concept of avoidable mortality makes it possible to distinguish between improvements in health care and the impact of policies outside the health sector that also impact on the public's health, such as tobacco and alcohol policies (Albert et al. 1996; Nolte et al. 2002).

In summary, the concept of avoidable mortality has limitations but provides a potentially useful indicator of health-care system performance. However, it is important to stress that high levels should not be taken as definitive evidence of ineffective health care but rather as an indicator of potential weaknesses that require further investigation. The next section explores the tracer concept – a promising approach that allows more detailed analysis of a health system's apparent suboptimal performance.

Tracer concept

The Institute of Medicine (IoM) in the United States proposed the concept of tracer conditions in the late 1960s as a means to evaluate health policies (Kessner et al. 1973). The premise is that tracking a few carefully selected health problems can provide a means to identify the strengths and weaknesses of a health-care system and thereby assess its quality.

Kessner et al. (1973) defined six criteria to define health problems appropriate for application as tracers. They should have: (i) a definitive functional impact, i.e. require treatment, with inappropriate or absent treatment resulting in functional impairment; (ii) a prevalence high enough to permit collection of adequate data; (iii) a natural history which varies with the utilization and effectiveness of health care; (iv) techniques of medical management which are well-defined for at least one of the following: prevention, diagnosis, treatment, rehabilitation; and (v) be relatively well-defined and easy to diagnose, with (vi) a known epidemiology.

The original concept envisaged the use of tracers as a means to evaluate discrete health service organizations or individual health care. Developed further, it might also be used at the system level by identifying conditions that capture the performance of certain elements of the health system. This approach would not seek to assess the quality of care per se but rather to profile the system's response to the tracer condition and aid understanding of the strengths and weaknesses of that system. By allowing a higher level of analysis such an approach has the potential to overcome some of the limitations of the cruder comparative studies outlined earlier.

The selection of health problems suitable for the tracer concept will depend on the specific health system features targeted. Thus, vaccine-

preventable diseases such as measles might be chosen as an indicator for public health policies in a given system. Measles remains an important preventable health problem in several European countries, as illustrated by continuing outbreaks and epidemics (WHO Regional Office for Europe 2003). This is largely because of inadequate routine coverage in many parts of Europe, despite the easy availability of vaccination. These problems persist despite successes in reducing measles incidence to below one case per 100 000 in most EU Member States except Greece (1.1/100 000), Malta (1.5/100 000), Ireland (2.3/100 000) and Romania (23.2/100 000) (WHO Regional Office for Europe 2007).

Neonatal mortality has been suggested as a possible measure for assessing access to health care. For example, there were substantial declines in birthweight-specific neonatal mortality in the Czech Republic and the former East Germany following the political transition in the 1990s (Koupilová et al. 1998; Nolte et al. 2000). Thus, in east Germany neonatal mortality fell markedly (by over 30%) between 1991 and 1996 due to improvements in survival, particularly among infants with low and very low birth weight (<1500 g) (Nolte et al. 2000). This has been attributed, in part, to reform of the system of health care after unification which increased the availability of modern equipment and drugs for high-quality neonatal care. As with perinatal mortality, international comparisons of neonatal mortality can be problematic. However, temporal comparisons within a given country can provide important insights into potential weaknesses or advances in the quality of neonatal care, as demonstrated in east Germany.

Other work has examined the use of diabetes as a measure of health system performance in relation to chronic illness (Nolte et al. 2006). Deaths from diabetes among young people have been interpreted as 'sentinel health events' that should raise questions about the quality of health-care delivery (McColl & Gulliford 1993). The optimal management of diabetes requires coordinated inputs from a wide range of health professionals; access to essential medicines and monitoring; and, ideally, a system that promotes patient empowerment. Measures of diabetes outcome may therefore provide important insights into primary and specialist care and their systems of communication.

Nolte and colleagues (2006) generated a measure of 'case-fatality' among young people with diabetes, using published data on diabetes incidence among young people for the period 1990–1994 and mortality

under the age of 40 years for the period 1994–1998 in twenty-nine countries. This mortality-to-incidence ratio varied more than ten-fold across countries, consistent with findings of cohort studies of mortality among young people with type I diabetes. The mortality-to-incidence ratio for diabetes thus appears to provide a means of differentiating countries' quality of care for people with diabetes. While solely an indicator of potential problems, this can stimulate more detailed assessments of the problems raised and what can be done to address them. Chapter 4.2 (Chronic care) explores this in more detail.

The way ahead

A large body of work aims to define how best to analyse health system performance at the level of population health and the multiplicity of strategies and approaches employed. This demonstrates that there is no definitive solution for this central challenge of managing a health (care) system effectively. The main goals of a health system can be defined easily but it is more difficult to identify a way of assessing whether these goals are being achieved and the extent to which apparent progress can be attributed to the health system or to other factors.

The overview presented here illustrates the conceptual underpinning of different measures in use, the information they provide and their major problems. General indicators of population health (for example, total and infant mortality, life expectancy, DALYs) remain important and provide tools that allow quick and simple assessments of total societal health system performance. Careful age- and sex-specific demographic measures of mortality over time can be strongly suggestive but generally such indicators provide only limited insights into specific components of the health-care system that impact on health. In contrast, more specific indictors of population health, such as cancer survival, give more detailed insights into discrete aspects of the health-care system but when used in isolation do not reveal information on other areas of the system that may be equally important. Also, single indicators often identify only the need for more in-depth investigation of process.

In conclusion, assessments of health system performance require a set of probes in order to examine different levels. Given the variation of settings within and between countries it is equally clear that there will be no single best combination. The range and balance across

levels will differ according to the context within which each system sits; the expectations and norms of those who assess performance; and whether longitudinal (within system) or cross-sectional (across populations) comparisons are employed. Of necessity, the combination will also vary according to the availability of appropriate data and the resultant limitations of those data.

Despite its many limitations, the concept of avoidable mortality remains a valuable indicator of health-care system performance. However, it is important to reiterate that the underlying concept should not be mistaken as definitive evidence of differences in the effectiveness of health care. Avoidable mortality should be interpreted as an indicator of potential weaknesses in health care that may require further investigation.

References

Albert, X. Bayo, A. Alfonso, J. Cortina, P. Corella, D (1996). 'The effectiveness of health systems in influencing avoidable mortality: a study in Valencia, Spain, 1975–90.' *Journal of Epidemiology and Community Health,* 50(3): 320-325.

Alvarez-Dardet, C. Ruiz, M (1993). 'Thomas McKeown and Archibald Cochrane: a journey through the diffusion of their ideas.' *British Medical Journal,* 306(6887): 1252–1255.

Arah, O. Westert, G. Hurst, J. Klazinga, N (2006). 'A conceptual framework for the OECD Health Care Quality Indicators Project.' *International Journal for Quality in Health Care,* 18 (Suppl. 1): 5–13.

Babazono, A. Hillman, A (1994). 'A comparison of international health outcomes and health care spending.' *International Journal of Technology Assessment in Health Care,* 10(3): 376–381.

Beaglehole, R (1986). 'Medical management and the decline in mortality from coronary heart disease.' *British Medical Journal,* 292(6512): 33–35.

Becker, N. Boyle, P (1997). 'Decline in mortality from testicular cancer in West Germany after reunification.' *Lancet,* 350(9079): 744.

Bennett, K. Kabir, Z. Unal, B. Shelley, E. Critchley, J. Perry, I. Feely, J. Capewell, S (2006). 'Explaining the recent decrease in coronary heart disease mortality rates in Ireland, 1985–2000.' *Journal of Epidemiology and Community Health,* 60(4): 322–327.

Berrino, F. De Angelis, R. Sant, M. Rosso, S. Bielska-Lasota, M. Coebergh, J. Santaquilani, M. & EUROCARE Working Group (2007). 'Survival for eight major cancers and all cancers combined for European adults

diagnosed in 1995–99: results of the EUROCARE-4 study.' *Lancet Oncology,* 8(9): 773–783.

Berrino, F. Gatta, G. Sant, M. Capocaccia, R (2001). 'The EUROCARE study of survival of cancer patients in Europe: aims, current status, strengths and weaknesses.' *European Journal of Cancer,* 37(6): 673–677.

Bots, M. Grobee, D (1996). 'Decline of coronary heart disease mortality in the Netherlands from 1978 to 1985: contribution of medical care and changes over time in presence of major cardiovascular risk factors.' *Journal of Cardiovascular Risk,* 3(3): 271–276.

Brenner, G. Altenhofen, L. Bogumil, W. Heuer, J. Kerek-Bodden, H. Koch, H (2000). *Gesundheitszustand und ambulante medizinische Versorgung der Bevölkerung in Deutschland im Ost-West-Vergleich*. Cologne: Zentralinstitut für die kassenärztliche Versorgung.

Britton, A. McKee M. Black, N. McPherson, K. Sanderson, C. Bain, C (1999). 'Threats to applicability of randomised trials: exclusions and selective participation.' *Journal of Health Services Research and Policy,* 4(2): 112–121.

Buck, D. Eastwood, A. Smith, P (1999). 'Can we measure the social importance of health care?' *International Journal of Technology Assessment in Health Care,* 15(1): 89–107.

Bunker, J (1995). 'Medicine matters after all.' *Journal of the Royal College of Physicians,* 29(2): 105–112.

Bunker, J. Frazier, H. Mosteller, F (1994). 'Improving health: measuring effects of medical care.' *Milbank Memorial Fund Quarterly,* 72(2): 225–258.

Butler, C. Currie, G. Anderson, W (2005). 'Do differences in data reporting contribute to variation in lung cancer survival?' *Journal of the National Cancer Institute,* 97(18): 1385.

Capewell, S. Beaglehole, R. Seddon, M. McMurray, J (2000). 'Explanation for the decline in coronary heart disease mortality rates in Auckland, New Zealand, between 1982 and 1993.' *Circulation,* 102(13): 1511–1516.

Capewell, S. Morrison, C. McMurray, J (1999). 'Contribution of modern cardiovascular treatment and risk factor changes to the decline in coronary heart disease mortality in Scotland between 1975 and 1994.' *Heart,* 81(4): 380–386.

Charlton J. Hartley, R. Silver, R. Holland, W (1983). 'Geographical variation in mortality from conditions amenable to medical intervention in England and Wales.' *Lancet,* 1(8326 Pt 1): 691–696.

Ciccolallo, L. Capocaccia, R. Coleman, M. Berrino, F. Coebergh, J. Damhuis, R. Faivre, J. Martinez-Garcia, C. Moller, H. Ponz de Leon,

M. Launoy, G. Raverdy, N. Williams, E. Gatta, G (2005). 'Survival differences between European and US patients with colorectal cancer: role of stage at diagnosis and surgery.' *Gut,* 54(2): 268–273.

Cochrane, A. Leger, A. Moore, F (1978). 'Health service 'input' and mortality 'output' in developed countries.' *Journal of Epidemiology and Community Health,* 32(3): 200–205.

Coleman, M. Babb, P. Damiecki, P. Grosclaude, P. Honjo, S. Jones, J. Knerer, G. Pitard, A. Quinn, M. Sloggett, A. De Stavola, B (1999). *Cancer survival trends in England and Wales 1971–1995: deprivation and NHS region*. London: The Stationary Office.

Colgrove, J (2002). 'The McKeown thesis: a historical controversy and its enduring influence.' *American Journal of Public Health,* 92(5): 725–729.

Collins, J (1982). 'The contribution of medical measures to the decline of mortality from respiratory tuberculosis: an age-period-cohort model.' *Demography,* 19(3): 409–427.

Cremieux, PY. Ouellette, P. Pilon, C (1999). 'Health care spending as determinants of health outcomes.' *Health Economics,* 8(7): 627–639.

De Leeuw, R. Cuttini, N. Nadai, M. Berbik, I. Hansen, G. Kucinskas, A. Lenoir, S. Levin, A. Persson, J. Rebagliato, M. Reid, M. Schroell, M. De Vonderweid, U. EURONIC Study Group (2000). 'Treatment choices for extremely preterm infants: an international perspective.' *Journal of Pediatrics,* 137(5): 608–616.

Elola, J. Daponte, A. Navarro, V (1995). 'Health indicators and the organization of health care systems in western Europe.' *American Journal of Public Health,* 85(10): 1397–1401.

Etches, V. Frank, J. Di Ruggiero, E. Manuel, D (2006). 'Measuring population health: a review of indicators.' *Annual Review of Public Health,* 27: 29–55.

European Commission (2007). *The European Health Survey System*. Brussels: European Commission (http://ec.europa.eu/health/ph_information/dissemination/reporting/ehss_en.htm).

Fairchild, A. Oppenheimer, G (1998). 'Public health nihilism vs pragmatism: history, politics, and the control of tuberculosis.' *American Journal of Public Health,* 88(7): 1105–1117.

Fenton, A. Field, D. Mason, E. Clarke, M (1992). 'Attitudes to viability of preterm infants and their effect on figures for perinatal mortality.' *British Medical Journal,* 300(6722): 434–436.

Ford, E. Ajani, U. Croft, J. Critchley, J. Labarthe, D. Kottke, T. Giles, W. Capewell, S (2007). 'Explaining the decrease in US deaths from coronary disease, 1980–2000.' *New England Journal of Medicine,* 356(23): 2388–2398.

Garne, E (2001). 'Perinatal mortality rates can no longer be used for comparing quality of perinatal health services between countries.' *Paediatric and Perinatal Epidemiology,* 15(3): 315–316.

Garne, E., Berghold, A. Johnson, Z. Stoll, C (2001). 'Different policies on prenatal screening programmes and induced abortions explain regional variations in infant mortality with congenital malformations.' *Fetal Diagnosis and Therapy,* 16(3): 153–157.

Gatta, G. Capoccacia, R. Coleman, M. Ries, L. Hakulinen, T. Micheli, A. Sant, M. Verdecchia, A. Berrino, F (2000). 'Toward a comparison of survival in American and European cancer patients.' *Cancer,* 89(4): 893–900.

Gravelle, H. Blackhouse, M (1987). 'International cross-section analysis of the determination of mortality.' *Social Science and Medicine,* 25(5): 427–441.

Holland, W (1986). 'The 'avoidable death' guide to Europe.' *Health Policy,* 6(2): 115–117.

Holland, W (1988). *European Community atlas of 'avoidable death'.* Oxford: Oxford University Press.

Hunink, M. Goldman, L. Tosteson, A. Mittelman, M. Goldman, P. Williams, L. Tsevat, J. Weinstein, M (1997). 'The recent decline in mortality from coronary heart disease, 1980–1990.' *Journal of the American Medical Association,* 277(7): 535–542.

Hussey, PS. Anderson, GF. Osborn, R. Feek, C. McLaughlin, V. Millar, J. Epstein, A (2004). 'How does the quality of care compare in five countries?' *Health Affairs (Millwood),* 23(3): 89–99.

Jack, R. Gulliford, M. Ferguson, J. Moller, H (2003). 'Geographical inequalities in lung cancer management and survival in south east England: evidence of variation in access to oncology services?' *British Journal of Cancer,* 88(7): 1025–1031.

Johansson, S (2005). 'Commentary: the pitfalls of policy history. Writing the past to change the present.' *International Journal of Epidemiology,* 34(3): 526–529.

Jougla, E. Papoz, L. Balkau, B. Maguin, P. Hatton, F (1992). 'Death certificate coding practices related to diabetes in European countries – the 'EURODIAB Subarea C' study.' *International Journal of Epidemiology,* 21(2): 343–351.

Jougla, E. Rossolin, F. Niyonsenga, A. Chappert, J-L. Johansson, L. Pavillon, G (2001). *Comparability and quality improvement of the European causes of death statistics.* Paris: INSERM.

Kelley, E. Hurst, J (2006). *Health care quality indicators project. Conceptual framework paper.* Paris: Organisation for Economic Co-operation and Development (OECD Health Working Papers No. 23).

Kelson, M. Farebrother, M (1987). 'The effect of inaccuracies in death certification and coding practices in the European Economic Community (EEC) on international cancer mortality statistics.' *International Journal of Epidemiology,* 16(3): 411–414.

Kessner, D. Kalk, C. Singer, J (1973). 'Assessing health quality – the case for tracers.' *New England Journal of Medicine,* 288(4): 189–194.

Kesteloot, H. Sans, S. Kromhaut, D (2006). 'Dynamics of cardiovascular and all-cause mortality in western and eastern Europe between 1970 and 2000.' *European Heart Journal,* 27(1): 107–113.

Kim, K. Moody, P (1992). 'More resources better health? A cross-national perspective.' *Social Science and Medicine,* 34(8): 837–842.

Kobayashi, S (2004). 'What caused the decline in breast cancer mortality in the United Kingdom?' *Breast Cancer,* 11(2): 156–159.

Koupilová, I. McKee, M. Holcik, J (1998). 'Neonatal mortality in the Czech Republic during the transition.' *Health Policy,* 46(1): 43–52.

Kricker, A (2002). *Ovarian cancer in Australian women*. Camperdown, NSW: National Breast Cancer Center.

Laatikainen, T. Critchley, J. Vartiainen, E. Salomaa, V. Ketonen, M. Capewell, S (2005). 'Explaining the decline in coronary heart disease mortality in Finland between 1982 and 1997.' *American Journal of Epidemiology,* 162(8): 764–773.

Lalonde, M (1974). *A new perspective on the health of Canadians*. Ottawa: Government of Canada.

Leon, D. Vagerö, D. Olausson, P (1992). 'Social class differences in infant mortality in Sweden: comparison with England and Wales.' *British Medical Journal,* 305(6855): 687–691.

Lopez, AD. Mathers, CD. Ezzati, M. Jamison, DT. Murray, CJL (2006). *Global burden of disease and risk factors*. New York: Oxford University Press and World Bank.

Mackenbach, J (1996). 'The contribution of medical care to mortality decline: McKeown revisited.' *Journal of Clinical Epidemiology,* 49(11): 1207–1213.

Mackenbach, J. Looman, C. Kunst, A. Habbema, J. van der Maas, P (1988). 'Post-1950 mortality trends and medical care: gains in life expectancy due to declines in mortality from conditions amenable to medical intervention in the Netherlands.' *Social Science and Medicine,* 27(9): 889–594.

Mackenbach, J. Van Duyne, W. Kelson, M (1987). 'Certification and coding of two underlying causes of death in the Netherlands and other countries of the European Community.' *Journal of Epidemiology and Community Health,* 41(2): 156–160.

Mariotto, A. Capocaccia, R. Verdecchia, A. Micheli, A. Feuer, E. Pickle, L. Clegg, L (2002). 'Projecting SEER cancer survival rates to the US: an ecological regression approach.' *Cancer Causes Control,* 13(2): 101–111.

Marques-Vidal, P. Ferrieres, J. Metzger, M. Cambou, J. Filipiak, B. Löwel, H. Keil, U (1997). 'Trends in coronary heart disease morbidity and mortality and acute coronary care and case fatality from 1985–1989 in southern Germany and south-western France.' *European Heart Journal,* 18(5): 816–821.

Mathers, C. Ma Fat, D. Inoue, M. Rao, C. Lopez, A (2005). 'Counting the dead and what they died from: an assessment of the global status of cause of death data.' *Bulletin of the World Health Organization,* 83(3): 171–77.

McColl, AJ. Gulliford, MC (1993). *Population health outcome indicators for the NHS. A feasibility study.* London: Faculty of Public Health Medicine and the Department of Public Health Medicine, United Medical and Dental Schools of Guy's and St Thomas' Hospitals.

McDowell, I. Spasoff, R. Kristjansson, B (2004). 'On the classification of population health measurements.' *American Journal of Public Health,* 94(3): 388–393.

McHaffie, H. Cuttini, M. Brolz-Voit, G. Randag, L. Mousty, R. Duguet, A. Wennergren, B. Benciolini, P (1999). 'Withholding/withdrawing treatment from neonates: legislation and official guidelines across Europe.' *Journal of Medical Ethics,* 25(6): 440–446.

McKee, M (1999). 'For debate – Does health care save lives?' *Croatian Medical Journal,* 40(2): 123–128.

McKee, M. Shkolnikov, V. & Leon, D (2001). 'Alcohol is implicated in the fluctuations in cardiovascular disease in Russia since the 1980s.' *Annals of Epidemiology,* 11(1): 1–6.

McKeown, T (1971). Medical issues in historical demography. In: Clarke, E (ed.). *Modern methods in the history of medicine.* London: Athlone Press.

McKeown, T (1979). *The role of medicine: dream, mirage or nemesis?* Oxford: Blackwell.

McKinlay, J. McKinlay, S (1977). 'The questionable contribution of medical measures to the decline of mortality in the United States in the twentieth century.' *Milbank Memorial Fund Quarterly,* 55(3): 405–428.

Murray, C. Salomon, J. Mathers, C (2000). 'A critical examination of summary measures of population health.' *Bulletin of the World Health Organization,* 78(8): 981–994.

Murray, C. Salomon, J. Mathers, C. Lopez, A (eds.) (2002). *Summary measures of population health. Concepts, ethics, measurement and applications.* Geneva: World Health Organization.

Nixon, J. Ulmann, P (2006). 'The relationship between health care expenditure and health outcomes. Evidence and caveats for a causal link. *European Journal of Health Economics,* 7(1): 7–18.

Nolte, E. Bain, C. McKee, M (2006). 'Chronic diseases as tracer conditions in international benchmarking of health systems: the example of diabetes.' *Diabetes Care,* 29: 1007–1011.

Nolte, E. Brand, A. Koupilova, I. McKee, M (2000). 'Neonatal and postneonatal mortality in Germany since unification.' *Journal of Epidemiology and Community Health,* 54(2): 84–90.

Nolte, E. McKee, M (2004). *Does healthcare save lives? Avoidable mortality revisited.* London: The Nuffield Trust.

Nolte, E. McKee, M (2008). 'Measuring the health of nations: updating an earlier analysis.' *Health Affairs,* 27(1): 58–71.

Nolte, E. McKee, M. Wait, S (2005). Describing and evaluating health systems. In: Bowling, A. Ebrahim, S (eds.) *Handbook of health research methods: investigation, measurement and analysis.* Maidenhead: Open University Press.

Nolte, E. Scholz, R. Shkolnikov, V. McKee, M (2002). 'The contribution of medical care to changing life expectancy in Germany and Poland.' *Social Science and Medicine,* 55(11): 1907–1923.

OECD (2007). *OECD health data 2007.* Paris: Organisation for Economic Co-operation and Development.

Or, Z (2000). 'Determinants of health outcomes in industrialised countries: a pooled, cross-country, time-series analysis.' *OECD Economic Studies,* 30: 53–77.

Or, Z (2001). *Exploring the effects of health care on mortality across OECD countries*. Paris: Organisation for Economic Co-operation and Development (Labour Market and Social Policy Occasional Paper No. 46).

Pennington, C (1979). 'Mortality and medical care in nineteenth-century Glasgow.' *Medical History,* 23(4): 442–450.

Reidpath, D (2005). 'Population health. More than the sum of the parts?' *Journal of Epidemiology and Community Health,* 59(10): 877–880.

Richardus, J. Graafmans, W. Verloove-Vanhorick, S. Mackenbach, J (1998). 'The perinatal mortality rate as an indicator of quality of care in international comparisons.' *Medical Care,* 36(1): 54–66.

Richardus, J. Graafmans, W. Verloove-Vanhorick, S. Mackenbach, J. EuroNatal International Audit Panel, EuroNatal Working Group (2003). 'Differences in perinatal mortality and suboptimal care between 10 European regions: results of an international audit.' *British Journal of Obstetrics and Gynaecology,* 110(2): 97–105.

Rutstein, D. Berenberg, W. Chalmers, T. Child, C. Fishman, A. Perrin, E (1976). 'Measuring the quality of medical care. A clinical method.' *New England Journal of Medicine,* 294(11): 582–588.

Ruzicka, L. Lopez, A (1990). 'The use of cause-of-death statistics for health situation assessment: national and international experiences.' *World Health Statistics Quarterly,* 43(4): 249–258.

Sant, M. Allemani, C. Berrino, F. Coleman, M. Aareleid, T. Chaplain, G. Coebergh, J. Colonna, M. Crosignani, P. Danzon, A. Federico, M. Gafà, L. Grosclaude, P. Hédelin, G. Macè-Lesech, J. Garcia, C. Møller, H. Paci, E. Raverdy, N. Tretarre, B. Williams, E. European Concerted Action on Survival and Care of Cancer Patients (EUROCARE) Working Group (2004). 'Breast carcinoma survival in Europe and the United States.' *Cancer,* 100(4): 715–722.

Schneyder, S. Landefeld, J. Sandiffer, F (1981). 'Biomedical research and illness: 1900–1979.' *Milbank Memorial Fund Quarterly,* 59(1): 44–58.

Scioscia, M. Vimercati, A. Maiorano, A. Depalo, R. Selvaggi, L (2007). 'A critical analysis on Italian perinatal mortality in a 50-year span.' *European Journal of Obstetrics, Gynecology, and Reproductive Biology,* 130(1): 60–65.

Szreter, S (1988). 'The importance of social interventions in Britain's mortality decline.' *Social History of Medicine,* 1(1): 1–37.

Tunstall-Pedoe, H. Kuulasmaa, K. Mahonen, M. Tolonen, H. Ruokokoski, E. Amouyel, P (1999). 'Contribution of trends in survival and coronary-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA project populations.' *Lancet,* 353(9164): 1547–1557.

Tunstall-Pedoe, H. Vanuzzo, D. Hobbs, M. Mähönen, M. Cepatis, Z. Kuulasmaa, K. Keil, U (2000). 'Estimation of contribution of changes in coronary care to improving survival, event rates, and coronary heart disease mortality across the WHO MONICA project populations.' *Lancet,* 355(9205): 688–700.

Unal, B. Critchley, J. Capewell, S (2007). 'Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000.' *Circulation,* 109(9): 1101–1107.

Uuskula, M. Lamp, K. Vali, M (1998). 'An age-related difference in the ratio of sudden coronary death over acute myocardial infarction in Estonian males.' *Journal of Clinical Epidemiology,* 51(7): 577–580.

Van der Pal-de Bruin, K. Graafmans, W. Biermans, M. Richardus, J. Zijlstra, A. Reefhuis, J. Mackenbach, J. Verloove-Vanhorick, S (2002). 'The influence of prenatal screening and termination of pregnancy on perinatal mortality rates.' *Prenatal Diagnosis,* 22(11): 966–972.

Van der Zee, J. Kroneman, M (2007). 'Bismarck or Beveridge: a beauty contest between dinosaurs.' *BMC Health Services Research,* 7(1): 94.

Verdecchia, A. Francisci, S. Brenner, H. Gatta, G. Micheli, A. Mangone, L. Kunkler, I. EUROCARE-4 Working Group (2007). 'Recent cancer survival in Europe: a 2000–02 period analysis of EUROCARE-4 data.' *Lancet Oncology,* 8(9): 784–796.

WHO (2000). *The world health report 2000. Health systems: improving performance*. Geneva: World Health Organization.

WHO Regional Office for Europe (2003). *Strategic plan for measles and congenital rubella infection in the European region of WHO*. Copenhagen: WHO Regional Office for Europe.

WHO Regional Office for Europe (2007). European Health for All database [offline database]. Copenhagen: WHO Regional Office for Europe (January update).

Wiesner, G. Grimm, J. Bittner, E (1999). 'Zum Herzinfarktgeschehen in der Bundesrepublik Deutschland: Prävalenz, Inzidenz, Trend, Ost-West-Vergleich.' *Gesundheitswesen,* 61(Suppl. 2): 72–78.

Wilson, L (2005). 'Commentary: medicine, population, and tuberculosis.' *International Journal of Epidemiology,* 34(3): 521–524.

Wright, J. Weinstein, M (1998). 'Gains in life expectancy from medical interventions – standardizing data on outcomes.' *New England Journal of Medicine,* 339(6): 380–386.

Zatonski, W. McMichael, A. Powles, J (1998). 'Ecological study of reasons for the sharp decline in mortality from ischaemic heart disease in Poland since 1991.' *British Medical Journal,* 316(7137): 1047–1051.

Annex 1 Summary measures of population health

Recent decades have seen a growing interest in, and work on, indicators that combine information on mortality and non-fatal health outcomes to summarize population health. Etches et al. (2006) distinguish two general categories of summary measures of population health: (i) health expectancies; and (ii) health gaps. Health expectancies determine how long people can expect to live free of certain diseases or limitations on their normal activities. In contrast, health gaps measure the difference between a specified health norm for the population (e.g. seventy-five as the average age at death) and the actual health of the population. The latter is most commonly assessed using DALYs.

Key issues include how to define and measure disability and select the weights to apply to particular health states. Disability weighting means that conditions which are disabling but rarely cause death (particularly mental illness) are ranked as more important than they would be if ranked by mortality alone. This is related to the highly controversial debate on the value placed on a year of life at different stages. For example, The Global Burden of Disease project (Lopez et al. 2006) placed more weight on a year of life of a young adult than on that of a child. This has the effect of reducing the burden of disease arising from deaths in childhood. One further issue concerns how to obtain estimates for countries from which data are unavailable. At present, these are often modelled on the relationships between mortality and other variables in countries which have data available. Given all of these issues, it is important to note that continuing debate surrounds the use of measures such as DALYs in policy-making.

Annex 2 Overview of selected measures of population health in health system performance assessment

Indicator	Definition	Advantages	Limitations
Life expectancy	Summary measure of the probability of dying at different ages in a population in a given year	Easy to understand and calculate as underlying mortality data are available for high- and most middle-income countries	Summarizes total mortality experience in a given population and therefore cannot be linked directly to activities within the health system
Age-standardized mortality rates by cause	Number of deaths per 100 000 population from a given cause adjusted for differences in the age-distribution between populations or over time	Easy to understand and calculate as underlying mortality data are available for high- and most middle-income countries	Cross-national comparisons might be limited by differences in cause of death certification in different settings and completeness of data registration
Infant mortality rate* (IMR)	Number of deaths in children within the first year of life, per 1000 live births in a given year	Easy to understand and calculate as underlying mortality data are available for high- and most middle-income countries	Combines neonatal and postneonatal deaths which are differentially sensitive to health-care quality (see text)

*Under-five mortality rate is also widely used, although mainly in low-income settings. This has the same structure but extends capture of deaths to five years so it is weighted more heavily than IMR towards influences of nutrition and primary care.

Perinatal mortality rate	Number of foetal deaths and deaths in the first week of life (early neonatal death) per 1000 live and stillbirths in a given year	Commonly used measure of quality of high-level care	Many challenges to interpretation, especially in international comparisons (see text)
Healthy life expectancy at birth (HALE)	Average number of years that a person can expect to live in 'full health' by taking into account years lived in less than full health because of disease and/or injury	Similar to life expectancy; takes account of the disease burden estimates available from the Global Burden of Disease study	Particular challenges in assigning acceptable disability weights
Disease-specific survival	Summary measure of the average length of time that individuals survive following diagnosis. It is most commonly used in respect to cancer.	Reasonably good comparative measure (but see text for limitations); data for cancer survival widely available	For cancers stage-specific data usually unavailable; inappropriate for evaluating screening programmes

2.2 *Patient-reported outcome measures and performance measurement*

RAY FITZPATRICK

Introduction

One of the most important developments in the assessment of health-care performance in recent years has been the demonstration that patients' and users' experiences of health and illness can be reliably and accurately captured by very simple means. It is now possible to capture aspects of health that are of most concern to individuals and populations – usually with self-completed and fairly short questionnaires. Typically these ask respondents to report, describe or assess aspects of their current health (e.g. symptoms); and the physical, psychological and social impact of health problems. The technical and scientific developments that have resulted in this capacity to capture patients' experiences have occurred over the last thirty years and these methods can now be considered mature, established and well-understood.

This chapter reviews the range of measures available and describes key considerations for selecting these for particular applications. It also considers the scope for widespread use of these measures to monitor health-care performance and the barriers that may limit such uses. Instruments in this field have been variously termed measures of quality of life, health status, health-related quality of life, subjective health status and functional status. The most important use of these questionnaires is for assessing outcomes of health care and increasingly they are referred to as patient-reported outcome measures (PROMs), the term used here.

Uses of PROMs

One of the simplest applications of PROMs is their use in surveys to assess the health of populations or segments of populations, e.g. users of particular facilities such as a hospital or clinic. For example,

the Health Survey for England (Joint Health Surveys Unit, 2008) is a household survey (usually of over 10 000 randomly selected adults) that gathers physiological and blood-sample based data and invites respondents to complete several questionnaire items about their health. The survey is conducted regularly and the information is an important resource to identify trends over time and geographical and social variations in health. Other more specialist national surveys are carried out from time to time to assess the prevalence and impact of disability in England and to assess the health of older people.

Increasingly, survey research to assess population levels of health is conducted on an international basis. For example, the Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national database of micro data on health, socio-economic status and social and family networks of individuals aged fifty or more, carried out across eleven European countries (Siegrist et al. 2007). Self-reported health is a major feature of this survey.

Health professionals also use PROMs in the context of individual patient care. Clinicians have argued that standard care in rheumatology is improved if, in addition to other clinical measures, PROMs are used regularly to assess a patient's current status with regard to pain and function (Pincus & Wolfe 2005). There are similar arguments that PROMs are essential to assess patients' needs and communication between patient and provider in routine care in other contexts such as oncology, dermatology and neurology (Lipscomb et al. 2007; Salek et al. 2007; Wagner et al. 1997).

In clinical trials PROMs can provide evidence that cannot be obtained by other means. This includes all the intended and unintended consequences of health-care interventions, whether drugs, new surgical techniques or innovations in the organization and delivery of services. In this sense they provide a necessary form of evidence of patient impact that complements the traditional clinical and laboratory measures employed. It is not yet standard practice to use PROMs in clinical trials (Sanders et al. 1998) but such uses have provided invaluable evidence of one key feature – they can provide evidence of change over time in an individual's health-related quality of life that can, in principle, be used as a means of assessing the performance or effectiveness of an intervention. Cross-sectional application of PROMs can be extended to longitudinal studies to offer a potential source of evidence of outcomes for determining health care's contribution to changes in health status.

PROMs are also being used more generally as evidence of outcomes to assess the contribution of health services to health in contexts such as professional quality assurance and audit and funders' assessments of the performance and value for money of services that they provide. Twenty years ago Ellwood (1988) proclaimed that PROMs offered a breakthrough for health services by providing funders and providers with evidence (for the first time) of benefits experienced by patients. It was argued that PROMs are uniquely important not only because they measure what matters to patients but also because they do so in ways that are feasible for large scale and regular use, such as through simple questionnaires. Claims are beginning to emerge, for example, in the Veterans Health Administration in the United States, that performance measurement (including PROMs) can be shown to improve the quality of care (Kerr & Fleming 2007).

Types of instruments

A bewildering number of PROMs exist. In 2002 my colleagues and I reported that our systematic review had identified at least 1275 such instruments in the English language alone (Garratt et al. 2002). We estimate that at least 3215 different instruments were reported in the English language literature in 2007.

PROMs can be classified into two basic types. *Generic* instruments have been developed to be relevant to the widest possible range of health problems. By contrast, *disease- or condition-specific* instruments are intended to be relevant to a limited disease or specific aspect or dimension of illness.

Generic instruments

Short-form 36 (SF-36) is by far the most commonly used generic measure (Ware & Sherbourne 1992). Thirty-six standard questions about the respondent's health in the last month are grouped into eight different dimensions of health: (i) physical functioning; (ii) role limitations due to physical problems; (iii) role limitations due to emotional problems; (iv) social functioning; (v) mental health; (vi) energy; (vii) pain; and (viii) health perceptions. As with most such instruments, responses are scored and all items in a given dimension (or scale) are combined to provide a single scale score, for example, for physical

functioning. Responses can also be used to produce just two, more general scale scores: a physical component summary and a mental component summary. Short-form 12 (SF-12) was developed as a shorter version that is normally scored to produce physical component summary and mental component summary scores. SF-36 has been translated into at least fifty languages and has been the object of more studies than any other instrument (Garratt et al. 2002). Its measurement properties (discussed in the following section) have been examined exhaustively, largely with very positive results.

Several other generic instruments have been widely used, notably the Sickness Impact Profile (Bergner et al. 1981), the Nottingham Health Profile (Hunt et al. 1985) and Dartmouth Primary Care Cooperative Information Project (Coop) Charts (Nelson et al. 1990). However, currently there is less supporting evidence for their use than for SF-36.

Utility instruments

In many ways utility instruments can be classed as generic instruments because they are all intended to have the widest applicability. However, unlike instruments such as SF-36, they were developed for one distinctive purpose – to assign overall values (or utilities) to respondents' health states. This overall value is particularly useful for analyses of the cost-effectiveness of health-care interventions. It allows researchers to estimate the overall aggregated value of the health states of the samples receiving an intervention, to allow comparisons of the costs. Traditional PROMs do not allow this overall calculation of the value of health states for individuals or aggregations of individuals.

EuroQol (EQ-5D) is the most commonly used utility instrument in Europe (Brooks 1996). This generic measure of health has five dimensions: (i) mobility; (ii) self-care; (iii) usual activities; (iv) pain/discomfort; and (v) anxiety/depression. Respondents choose between three levels of severity for each of the five dimensions and identify their position on a visual analogue scale ranging from zero (worst imaginable health state) to one hundred (best imaginable health state). A single weighted score (value) of the individual's health can be calculated from the five selected responses, using weights of values provided by a general population survey.

The Health Utilities Index (HUI) is the next most commonly used approach for deriving the values, preferences or utilities of respon-

dents. To date, less evidence is available to support the questionnaire-based versions of the HUI. Potential users have to decide between different interview formats and weigh the benefits of interview-based methods against the extra costs.

Disease-specific instruments

Disease-specific instruments have increased most rapidly in the last ten years (Garratt et al. 2002). They are developed to provide questionnaire content that is tailored to the specific disease for which they are intended. Thus, an instrument to assess health-related quality of life in rheumatoid arthritis might include specific items that would not feature in a generic instrument, e.g. on stiffness, fatigue or the difficulties of performing household tasks with hands. An instrument for Parkinson's disease might contain items concerning the consequences of tremor (e.g. holding objects, embarrassment in public) that would not be salient in a generic instrument. Typically disease-specific instruments are developed and explicitly identified as having relevance for an identified illness. The Arthritis Impact Measurement Scales and Parkinson's Disease Questionnaire (PDQ-39) have the specialist function conveyed by their titles.

The main reason for the growing interest in disease-specific PROMs is the belief that they are necessary to identify the small but important benefits and harms associated with novel interventions in clinical trials. This has some supporting evidence. Also, the broadly-based questionnaire content of generic instruments may miss both types of consequence.

Other instruments have been developed for more specific purposes, for example to assess outcomes in relation to specific health-care interventions. The Oxford knee score was developed specifically to assess outcomes of knee replacement surgery; a parallel PROM (Oxford hip score) is used to assess outcomes of hip replacement surgery. There is substantial evidence that these instruments are more sensitive to the specific problems of severe pain and the function of the patients receiving these procedures (Murray et al. 2007).

Individualized instruments

Recent years have seen the emergence of a number of instruments

based on a single important principle – individuals have their own personal goals and concerns in relation to health. Hence, traditional questionnaires with fixed items that are uniform for all respondents run contrary to the personal nature of health-related quality of life. Several new instruments attempt to elicit individuals' personal goals and concerns in a more flexible form. For example, the Patient Generated Index (Ruta et al. 1994) asks respondents to list the five most important areas of their lives that are affected by a disease or health problem; to rate how badly affected they are in each area; and to allocate points to the areas in which they would most value an improvement. Individual area ratings are weighted by the points given and summed to produce a single index. This is designed to measure the extent to which a patient's actual situation falls short of their hopes and expectations in those areas of life in which they would most value improvement. Such approaches are quite different to PROMs but the most obvious disadvantage of all the individualized instruments developed to date is the limited evidence for large-scale use. Generally, they require quite time-consuming and complex interviews.

Evaluating PROMs

A disciplined approach is needed to select an instrument for a particular application and formal criteria can inform the selection of sound choices from among the enormous number of instruments. Seven criteria are commonly used to assess PROMs: (i) reliability; (ii) validity; (iii) responsiveness; (iv) precision; (v) interpretability; (vi) acceptability; and (vii) feasibility (Fitzpatrick et al. 1998). It is possible to inspect the published evidence and weigh the amount of positive evidence for an instrument under each of these criteria. However, appropriateness is the eighth and arguably the most important criterion as it asks whether an instrument is relevant to the specific purpose of a given user. This requires judgements on (for example) the match between the content of the instrument and the purpose of the user. Such judgements are context specific and less easily informed by the general literature on an instrument.

Reliability is a fundamental requirement of any system of measurement. The more reliable an instrument, the more it is free of error. The literature is written as if reliability is a fixed feature of a PROM but it is dependent on the specific population in which it is used.

The reliability of PROMs is usually estimated in terms of internal consistency and reproducibility. It has already been pointed out that PROMs commonly take the form of scales (e.g. SF-36) – questionnaire items that combine to measure a construct such as pain or social isolation. The greater the agreement between the items of the scale the higher its internal consistency. This is one aspect of reliability and a variety of statistical tests have been developed to assess the extent to which scales reach perfect consistency. However, there is a complication to this approach. Perfect internal consistency is achieved most easily when questionnaire items are virtually identical to each other (i.e. asking the same question). Such an instrument is not desirable in practice. Instruments require scales that capture the different facets or aspects of, say, pain or social isolation. This is more likely to be achieved with items that do not correlate perfectly. As a result of these contrasting requirements, internal consistency statistics of instruments are expected to be high, but not too high (no higher than 0.90 on a range from 0.00 to 1.00).

Reproducibility is the other aspect of reliability. This the extent to which a measuring instrument produces the same result on repeated use, as long as the construct it is measuring does not change. A variety of test statistics have been developed to express the extent to which instruments are consistent over time. Typically instruments are retested on respondents between two and fourteen days after the first administration. An additional check to confirm that respondents have not experienced any change in their health (for example, using a simple global question) can be used to focus reproducibility estimates on stable respondents when assessing the reproducibility of a PROM.

Validity concerns the extent to which an instrument measures what it purports to measure. As with reliability, an instrument is only validated in the contexts in which its validity has been tested. Again, the literature generally overlooks this point but it is misleading to call an instrument validated without some qualification. Thus, an instrument validated to assess disability in multiple sclerosis may not be valid to assess disability in epilepsy as the measurement properties need to be re-established in the new context. The literature on validation of PROMs is dense and complex and only three key points are emphasized here.

Firstly, criterion validity assesses the extent to which scores from a new instrument agree with those of a gold standard. This has little

relevance as it is rare for a new instrument to be necessary or justified if a gold standard exists. Second, content validity is always crucial in judging a PROM, although it is a matter of judgement rather than statistical testing. Evidence of content validity is provided by clear explanation of what the instrument is intended to measure; how the items were developed and chosen; and whether these items appear to cover the intended construct. Third, the construct validity of an instrument is statistical. This assesses the available evidence in relation to the extent to which scores from the instrument agree with other measures in ways that are expected. Increasingly, authors of new PROMs are required to specify hypotheses of how they expect the test instrument to relate to other variables in order to avoid the biases of retrospective logic.

Responsiveness addresses the extent to which an instrument is able to detect changes over time in respondents' health. Since the overarching goal of health care is to bring about beneficial change, it may be argued that the most important requirement is that an instrument should accurately capture changes in health when they occur. Sometimes an instrument needs to detect clinically important changes (that is, not minor or 'noise'). However, it is argued increasingly that the term 'clinically' is unhelpful – changes have to be important and significant for the patient, not the health professional. A wide array of different statistical techniques is used to assess responsiveness, but no single approach dominates. The common thread is to assess the amount of variability in the change scores of an instrument that is due to change relative to other sources of variability (measurement error, patient characteristics, and so on).

Precision presents a problem for PROMs. This stems from the basic requirement to transform answers to questionnaires into quantitative scores that reflect accurately the full spectrum of the underlying phenomenon – pain, disability, social function, and so on. The following simplified example demonstrates how measurement assumptions may be problematic. An instrument with a physical mobility scale of, say, ten questionnaire items may be summed simply to produce a disability score. By accident of development, the majority of these items assess quite mild disability, for example, being unable to walk very long distances. An intervention that enabled a patient to improve at the mild end of the spectrum could produce improvements in the majority of items when assessed on the hypothetical scale. A patient with more disabil-

ity could improve at the severe end of the spectrum but show improvement on a smaller number of items. The latter result would be purely an artefact of questionnaire selection. An elegant study by Stucki and colleagues (1996), in which patients completed the physical mobility scale of SF-36 before and after hip-replacement surgery, shows that this is not just a hypothetical problem. Recently applied statistical techniques such as Rasch analysis are intended to address this problem by ensuring that scales for newly developed instruments provide unidimensional and interval-level measurement of domains (Norquist et al. 2004).

Interpretability is concerned with the meaning and inferences that may be drawn from an instrument. Typically, a PROM expresses changes that arise from a health-care intervention in terms of quantitative change scores on a scale that has little inherent meaning. It is possible to address the statistical significance of a given change score but less easy to give the result intuitive meaning. One approach is to equate a PROM's change scores to some other life event (if such evidence is available), for example to show that a change score is equivalent to the deterioration in health associated with a major life event such as bereavement. Another approach is to relate change scores to different levels of severity of illness, for example by comparing inpatients with less-severely ill patients in the community. Such approaches have not found much favour and it is likely that the field will increasingly resort to a different approach to identify minimal important differences for PROMs. This is outlined below.

Acceptability is an essential requirement. If respondents do not like a PROM they will either leave items incomplete or fail to answer the questionnaire at all, with major risk of bias in the interpretation of results. Instruments vary substantially in simple factors such as length and completion time. There are also less obvious variations, such as the amount of distressing or complex judgements required from the respondent.

Few studies directly address the issue of acceptability. One exception is a study of patients who were followed up after attendance at eighteen Swedish hospitals (Nilsson et al. 2007). Respondents were asked to complete both SF-36 and EQ-5D and to comment on their satisfaction with the two instruments. The majority appeared equally happy with both but a minority expressed a clear preference. Of these, more preferred SF-36 and commonly stated that it allowed them to

report their health more comprehensively or that the response categories allowed more nuanced answers. In another study (Moore et al. 2004), patients with multiple sclerosis received SF-36, EQ-5D and a fifty-four item disease-specific instrument. The majority were happy with all three instruments but the longer disease-specific instrument was preferred among the minority who expressed a preference. By way of contrast, patients in a follow-up to a major trial of treatment for stroke were randomized to report their health using either EQ-5D or SF-36 (Dorman et al. 1997). Respondents who received EQ-5D showed a higher response rate and fewer responses with incomplete data. However, acceptability may depend as much on specific features of the respondent group such as age, co-morbidity and the reason for involvement in a survey.

Feasibility needs to be considered separately as more resources are necessary for instruments that require trained staff or that involve significant transformation or processing of data to derive results. Costs become a major consideration if PROMs are to be delivered to large samples and/or over long periods of time.

Evidence to aid choice of instrument

It is clear that there is a burgeoning number of instruments from which to choose for any given problem and that evidence of their measurement properties and performance is potentially complex. It is not surprising that increasing attention is given to comparing instruments to identify those PROMs that have overall superior performance. It is rare to randomize respondents between instruments to compare performance as in the study cited above (Dorman et al. 1997). It is far more common for patients to be asked to complete two or more PROMs in the context of a trial and to compare their performance. This can be very informative if the trial provides other information about health as a benchmark. Such studies have tended to focus on the comparative evidence of the responsiveness of instruments since this is the most critical requirement for evaluations of interventions.

Several studies have shown that shorter instruments are as sensitive to change as longer instruments (Fitzpatrick et al. 1993; Katz et al. 1992). This is significant because it suggests that instruments may be shortened and reduce respondents' burden without loss of important information. Studies have reported statistically driven reductions of

longer instruments such as the Sickness Impact Profile in which the short-form versions appear to produce similar results (de Bruin et al. 1994). Moran et al. (2001) used simulation techniques and results of a dataset of three trials of respiratory rehabilitation to analyse the consequences of reducing items in the scales of the widely used Chronic Respiratory Questionnaire. They found modest losses of reliability, validity and responsiveness that became serious losses only when the number of items was reduced to one per scale. It is likely that content validity, the degree of coverage of the underlying construct, is adversely affected when a scale comprises only one item.

Coste et al. (1997) reviewed a series of forty-two studies that used shorter but equivalent instruments and identified a number of problems. For example, analysis of the shorter version from the dataset in which respondents had completed a longer version produced artificially elevated correlations between the shorter and longer versions. Studies seldom re-examined the content validity of the new, shorter instrument. Coste et al. concluded that a shortened PROM needs to be re-assessed as if it is a brand new instrument, distinct from the longer original.

Disease-specific versus generic instruments

Comparative studies have also investigated the widely debated issue of the relative merits of disease-specific and generic PROMs. The argument for disease-specific instruments is based on the belief that such measures will be more sensitive to changes in the health-related quality of life produced by an intervention, mainly because they contain a higher proportion of supposedly relevant items for the illness and intervention being studied. However, some studies have failed to identify such advantages empirically. Walsh et al. (2003) invited patients with various conditions that produce back pain to participate in a longitudinal survey of health-related quality of life involving the completion of both disease-specific and generic PROMs. They found no evidence that the disease-specific instrument was more sensitive to change over time.

Wiebe et al. (2003) carried out a structured review and identified forty-three randomized controlled trials which included direct comparison of disease-specific and generic PROMs completed by the same patients. The Sickness Impact Profile, Nottingham Health

Profile and SF-36 generic instruments were most commonly used in the sample of trials. No significant difference between the two types of instrument was found when the trials with modest and small overall effects were sub-divided according to the size of the underlying treatment effect. The difference between the two types of measure became greater and more significant as the true underlying therapeutic effect became greater in trials, with disease-specific instruments consistently more responsive. This evidence of the superior responsiveness of disease-specific PROMs is consistent with a review by Murawski and Miederhoff (1998) who used a wider array of observational, as well as randomized, studies. Wiebe et al. (2003) caution that such evidence does not prove that all disease-specific measures are more responsive than all generic measures.

Increasingly, it is becoming necessary to carry out reviews that assess all of the available evidence in order to inform choices between instruments. For practical, largely clinical, reasons such reviews tend to focus on the evidence on PROMs that relate to specific illnesses. Some of these reviews are relatively informal in terms of how evidence is sought, assessed and described (Carr et al. 1996). However, they are becoming more formal with increasing use of explicit search and inclusion criteria for relevant studies and scoring of the strength and quality of evidence for instruments included in the review (Garratt et al. 2004; Haywood et al. 2005). This enables readers to draw independent assessments of the evidence to determine whether or not they agree with reviewers' recommendations.

It is has been argued that such reviews are helpful in facilitating evidence-based recommendations but frequently are still limited by their reliance on informal and implicit criteria for what constitutes good measurement properties (Terwee et al. 2007). For example, a review may report and rate all the available evidence on the validity, responsiveness and interpretability of instruments. Typically, this will not spell out explicitly what counts as evidence of good validity or responsiveness. As an example, Terwee et al. (2007) suggest that reviewers might require at least 75% of the specific hypotheses spelt out in advance of a study to be supported as positive evidence for an instrument's construct validity. Evidence falling short of this standard would be rated either indeterminate or negative. Terwee et al. argue that wide application of such standards would make reviews even more transparent and offer easy choices. These standards would

drive up the quality of reporting in original studies that assess the measurement properties of PROMs as these tend to be vague about most details of procedure.

Broad problems remain with reviews of the comparative value and performance of PROMs. Firstly, these are heavily influenced as much by the volume of evidence as by its quality – instruments tend to be rated as relatively poor largely because of a lack of evidence. Secondly, even the most explicit reviews require difficult judgments of the relative importance of different criteria. For example, many would argue that content validity is fundamentally important and cannot be substituted by good evidence on other criteria. Those who use PROMs in evaluative research often tend to prioritize responsiveness as their most important criterion for evaluating and selecting the instrument. The third and related problem is the difficulty of reviewing the evidence for instruments against all possible uses in all contexts. Unavoidable elements of judgment remain regardless of the methodological thoroughness of reviews.

Barriers to implementation

Clearly, a substantial number of well-validated PROMs are available to provide important evidence of health from users' and the community's perspectives. Nevertheless, health-care providers do not use PROMs widely on a regular basis. A number of studies have examined potential barriers to more widespread implementation. These may be grouped into two broad categories: (i) cognitive; and (ii) logistic and resource factors. Evidence for each of these is examined in turn.

Cognitive barriers

Health professionals' attitudes to PROMs have had a major influence on implementation. This is particularly true among doctors who have been found to be generally sceptical about their value. An early and influential review (Deyo & Patrick 1989) argued that doctors' training leads them to be distrustful of data that they consider subjective and soft. Information from questionnaires is viewed as inherently less reliable than biologically derived data. A study of oncologists found that they considered quality of life to be a very important issue for their patients but preferred to collect data informally. They were not

convinced of the validity of PROMS outside clinical trials (Taylor et al. 1996). A study of UK psychiatrists found that few clinicians regularly used PROMs in their daily practice (Gilbody et al. 2002). Many respondents explained this infrequent use by expressing scepticism about the reliability, validity and responsiveness of the available instruments. McHorney and Bricker (2002) asked doctors in a primary-care setting about the value of PROMs when assessing patients' function. Doctors were sceptical that questionnaire-based information could add to what was established by traditional history taking. A related problem was found in a study of Dutch paediatricians (Baars et al. 2004). They acknowledged that PROMs could provide valuable information in principle but were concerned that they lacked the skills and professional background to interpret and use the information provided by such instruments.

These reservations relate to a broader set of concerns. PROMs are seen to be of doubtful value as they do not improve a doctor's ability to diagnose and treat problems more effectively; they identify problems that a doctor can do nothing about and are therefore not an effective resource or intervention. Certainly randomized controlled trials that evaluated PROMs as an addition to clinical services have tended to be disappointing. For example, Kazis et al. (1990) randomized doctors to receive or not receive information from disease-specific health status instruments completed by their patients with rheumatoid arthritis every three months for a year. Doctors who received this form of feedback found it useful. However, comparison with controls showed no differences in processes of care such as medication, referral or satisfaction and no differences in health status at one year follow-up.

An early structured review of studies that experimentally evaluated the benefits of PROMs for patient care and outcomes was unable to find clear evidence to support their use (Greenhalgh & Meadows 1999). A variety of reasons have been suggested for these predominantly negative results. It may be that it is not inherent problems of data from PROMs per se but rather that the details of the timing, presentation and feed-back of data to health professionals limit their impact in trials. PROMs have been of particular and long-standing interest in cancer services and some more encouraging and more focused studies have started to emerge in that field. Detmar et al. (2002) randomized doctors in a outpatient palliative care clinic to provide standard care alone or, with the addition of three consecutive

outpatient visits, in combination with graphic summaries of patients' quality of life recorded by a cancer-specific questionnaire. Audiotapes of consultations were analysed. Health-related quality of life was discussed significantly more frequently in the consultations for which doctors received patients' quality of life scores. Also, the experimental consultations identified a higher proportion of health problems than the controls.

A similarly positive result was obtained in a trial by Velikova et al. (2004). They randomly assigned patients to be either controls receiving usual care or in an experimental arm that involved regular completion of a cancer-specific PROM with results fed back to their doctors. At the end of the study the experimental group's consultations had more discussion of health-related quality of life and also experienced more favourable quality of life than the controls. The investigators noted greater improvement in health-related quality of life in those patients who had explicitly discussed the subject during consultations. In discussing the differences with other, negative, studies the investigators also note that their patients saw different clinicians sequentially. PROMs may be more valuable in these situations than where there is strong continuity of care. More encouraging evidence from more recent trials probably reflects more appropriate instruments and better ways of feeding information into clinicians' routines (Marshall et al. 2006).

The uptake of PROMs may also have been hindered by the belief that such questionnaires are intrusive or burdensome to patients and therefore jeopardize the professional-client relationship. Studies that have included a separate assessment of patients have invariably found that the majority consider that the information conveyed by their responses is important for health professionals to know and are positively satisfied with the task of completing a questionnaire (Detmar et al. 2002; McHorney & Bricker 2002; Nelson et al. 1990). In the study by McHorney et al. (2002), some patients queried the appropriateness of items on anxiety and depression in the context of what they considered to be purely physical presenting problems.

Logistic and resource barriers

Logistic and resource barriers include a set of related practical considerations. Time is one that immediately concerns health professionals.

In the study of oncologists and their views about PROMs described earlier (Taylor et al. 1996), 85% of respondents felt that time constraints made it difficult to integrate PROMs into routine patient care. Time was also the most commonly cited obstacle in the survey of paediatricians (Baars et al. 2004). The doctors in the study by McHorney and Bricker (2000) felt that the economics of managed care meant that there was no time for additional activities such as assessment of patients' answers to PROMs. The psychiatrists in the study by Gilbody et al. (2002) also felt that more time would be required to include PROMs in regular care.

Time is related to the broader challenge described in different ways in the various studies of the use of PROMs in routine practice – the need for significant changes in administrative routines in order to incorporate regular use of PROMs. Gilbody et al. reported that psychiatrists emphasized the need for a 'robust infrastructure, particularly in terms of administration and information technology resources' in order to incorporate the routine use of PROMs (Gilbody et al. 2002, p102). The American doctors would require the whole 'office ecosystem' to be changed (McHorney & Bricker 2002, page 1117). However, administrative changes are not enough. The basic routines of health professionals would require adjustments to enable PROMs to become a core part of a clinical service.

Economic costs are frequently cited as an additional consideration but it is remarkable how few attempts have been made to estimate such costs. Moinpour et al. (2007) were unable to provide any estimate of the costs of including PROMs in cancer trials because they were invariably bundled in with other research costs. They were able to conclude only that the costs of PROMs were likely to prove considerably lower than other clinical and biological endpoints. A recently published study by a group at the London School of Hygiene and Tropical Medicine (2007) provides one of the few explicitly calculated estimates of the total costs of collecting longitudinal data on PROMs. They conclude that the total costs in relation to elective surgical procedures are approximately £ 6.50 per patient included in a longitudinal survey. The majority of costs relate to data entry and they suggest that there may be significant opportunities for cost reduction. This issue will require further investigation if widespread use of PROMs is to be contemplated in health-care systems.

Current and future issues

One trend can confidently be predicted – continued proliferation of PROMs despite the attendant confusion that is risked by the availability of ever larger numbers of instruments. The pharmaceutical industry is likely to be the main driver of this growth as it responds to the growing need to demonstrate impact on ever more specific aspects of health-related quality of life for the burgeoning chronic disease market. Regulatory pressures in particular will drive the industry to use clearly validated instruments to demonstrate ever more precisely prespecified domains of quality of life in specific diseases.

The proliferation of instruments will be driven by the recognition that disease-specific instruments can be developed to incorporate the preference- or utility-based measurement required for health economic decisions (e.g. Torrance et al. 2004). It is not hard to foresee a plethora of instruments that produce increasingly difficult selection choices and growing problems with the non-comparability of the results of trials and evaluative studies that use increasingly different measures for similar domains of outcome. The capacity to provide reliable reviews and assessments of the quality and performance of the growing array of PROMs will need to be constantly improved.

It is often argued that trials and evaluative studies aiming to address health-related quality of life (particularly health) problems should include both a disease-specific and a generic measure in order optimally to capture the full spectrum of outcomes. It may also be argued that more short generic instruments such as EQ-5D or SF-12 are needed to complement disease-specific PROMs. They will provide some means of maintaining comparability of outcomes across studies given the increasing proliferation of disease-specific measures.

A potentially important development that is intended to solve many of the problems concerning the proliferation of PROMs may simply add to difficulties in the short term. The PROMIS initiative is sponsored by the National Institutes of Health in the United States. As discussed above it is a large-scale collaboration between scientists that will draw on existing instruments and develop new items (Cella et al. 2007a) for investigators to use in trials and evaluative studies. The long-term vision is to ensure that patients and populations will be

assessed by items that are maximally relevant to respondents' specific health problems and levels of disability. In some respects this vision resembles that driving the emergence of the individualized PROMs described earlier. However, PROMIS involves two quite new techniques to identify standard questionnaire items that maximally match the health of the respondent (Cella et al. 2007). Firstly, item response theory is a statistical method to select items that match respondents' levels of health or disability. Secondly, computerized adaptive testing uses the many strengths of information technology to facilitate that matching process. To provide a grossly simplified example – a respondent at a computer answers one question on health and is efficiently moved on to the next most appropriate question because the system takes account of the answer to the first question. Overall, the volume and redundancy of items required of the respondent is minimized and the assessment burden is reduced. PROMIS has only been in existence since 2005 so it is difficult to assess achievements. In the short term the very flexibility of such measuring systems may be confusing for potential users who are familiar with conventional, standard, fixed instruments.

Another potentially important recent development has been the publication of a document by the American Food and Drug Administration (FDA) (http://www.fda.gov/CDER/GUIDANCE/5460dft.pdf). This is likely to be widely influential as it describes in some detail how evidence from PROMs for drugs and medical products is assessed and underlines the importance of issues such as analysis of the implications of missing data for PROMs. Its most striking discussion concerns the need for those who use this evidence to have a very clearly developed model of how a product or drug might relate to quite specific aspects of health-related quality of life and to submit detailed evidence of a PROM's validity in measuring those specific domains. At the very least this will require much more careful consideration of the selection and justification of instruments for use in trials. Ritual inclusion of SF-36 or EQ-5D to address quality of life aspects in an unfocused way will no longer be a valid strategy, at least for submissions to the FDA.

These recent trends are emerging from the pharmaceutical industry and its regulators and push PROMs to become ever more specialized and targeted instruments. There will be greater need for health-care funders, providers and regulators to produce broader evidence of outputs and outcomes via PROMs but as yet this is not articulated as forcefully. It might be expected that these needs will push towards

more generic solutions that capture the broad impacts of services on patients and the public. It may be that the field increasingly diverges between these increasingly different needs of industry and public services. It will be a challenge for the science to respond to increasingly diverse expectations.

Policy implications

As yet, there is no evidence of PROMs being used extensively and routinely in a health-care system in order to assess performance and improve quality. The National Health Service (NHS) in the United Kingdom requires health-care providers to monitor four major elective surgical procedures (primary hip or knee replacement, groin hernia surgery, varicose vein procedures) by means of specified PROMs from 2009. This decision has enormous significance as it is the first real test of the scale of benefits that may accrue to patients, the public and providers when representative evidence from PROMs is available to assess the outcomes of all public service providers of particular interventions. It is significant that the decision to make monitoring of outcomes by PROMs effectively compulsory for four elective surgical procedures was preceded by structured reviews to identify the best performing PROMs for the four procedures. These were followed by pilot studies to ensure that the most appropriate PROMs could be identified and that it was feasible to use them for longitudinal monitoring. It is also significant that these four surgical procedures have a fairly clear, well-understood and specific role in relation to patients' health status. It will be interesting to see how readily the NHS moves from applying PROMs in the relatively simple environment of elective surgery to assessing the outcomes of long-term conditions for which the benefits of interventions may be less clear cut.

To date, PROMs' real world impact on routine services is largely theoretical and assumed. The NHS is field-testing the potential for PROMs to improve decisions about health care. The real challenge will be to examine their contribution to patients' and providers' decisions in relation to more complex health problems where multiple services over time make modest and often hard to define contributions to the quality of life. These contributions will need careful piloting and evaluation before services will feel confident to embrace PROMs on a widespread and regular basis.

References

Baars, R. van der Pal, S. Koopman, H. Wit, J (2004). 'Clinicians' perspective on quality of life assessment in paediatric clinical practice.' *Acta Paediatrica,* 93(10): 1356–1362.

Bergner, M. Bobbitt, R. Carter, W. Gilson, B (1981). 'The sickness impact profile: development and final revision of a health status measure.' *Medical Care,* 19(8): 787–805.

Brooks, R (1996). 'EuroQol: the current state of play.' *Health Policy*, 37(1): 53–72.

Carr, AJ. Thompson, PW. Kirwan, JR (1996). 'Quality of life measures.' *British Journal of Rheumatology*, 35(3): 275–281.

Cella, D. Gershon, R. Lai, J. Choi, S (2007). 'The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment.' *Quality of Life Research*, 16(Suppl. 1): 133–141.

Cella, D. Yount, S. Rothrock, N. Gershon, R. Cook, K. Reeve, B. Ader, D. Fries, JF. Bruce, B. Rose, M (2007a). 'The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years.' *Medical Care,* 45(Suppl. 1): 3–11.

Coste, J. Guillemin, F. Fermanian, J (1997). 'Methodological approaches to shortening composite assessment scales.' *Journal of Clinical Epidemiology,* 50(3): 247–252.

de Bruin, A. Diederiks, J. de Witte, L. Stevens, F. Phillipsen, H (1994). 'The development of a short generic version of the sickness impact profile.' *Journal of Clinical Epidemiology,* 47(4): 407–418.

Detmar, S. Muller, M. Schornagel, J. Wever, L. Aaronson, N (2002). 'Health-related quality of life assessments and patient physician communication: a randomized controlled trial.' *Journal of the American Medical Association,* 288(23): 3027–3034.

Deyo, R. Patrick, D (1989). 'Barriers to the use of health status measures in clinical investigation, patient care and policy research.' *Medical Care,* 27(Suppl. 3): 254–268.

Dorman, P. Slattery, J. Farrell, B. Dennis, M. Sandercock, P (1997). 'A randomised comparison of the EuroQol and short form-36 after stroke.' *British Medical Journal,* 315(7106): 461–462.

Ellwood, P (1988). 'Shattuck lecture – outcomes management. A technology of patient experience.' *New England Journal of Medicine*, 318(23): 1549–1556.

Fitzpatrick, R. Ziebland, S. Jenkinson, C. Mowat, A (1993). 'A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis.' *Journal of Rheumatology*, 20(3): 429–436.

Fitzpatrick, R. Davey, C. Buxton, M. Jones, D (1998). 'Evaluating patient-based outcome measures for use in clinical trials.' *Health Technology Assessment,* 2(14): 1–74.

Garratt, AM. Brealey, S. Gillespie, WJ. DAMASK Trial Team (2004). 'Patient-assessed health instruments for the knee: a structured review.' *Rheumatology*, 43(11): 1414–1423.

Garratt, A. Schmidt, L. Mackintosh, A. Fitzpatrick, R (2002). 'Quality of life measurement: bibliographic study of patient assessed health outcome measures.' *British Medical Journal,* 324(7351): 1417–1422.

Gilbody, S. House, A. Sheldon, T (2002). 'Psychiatrists in the UK do not use outcomes measures.' *British Journal of Psychiatry,* 180: 101–103.

Greenhalgh, J. Meadows, K (1999). 'The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review.' *Journal of Evaluation and Clinical Practice*, 5(4): 401–416.

Haywood, K. Garratt, AM. Fitzpatrick, R (2005). 'Quality of life in older people: a structured review of generic self-assessed health instruments.' *Quality of Life Research*, 14(7): 1651–1668.

Hunt, S. McEwan, J. McKenna, S (1985). 'Measuring health status: a new tool for clinicians and epidemiologists.' *Journal of Royal College of General Practitioners,* 35(273): 185–188.

Joint Health Surveys Unit (2008). *Health survey for England 2007: latest trends*. Leeds: NHS Information Centre (http://www.ic.nhs.uk).

Katz, J. Larson, M. Phillips, C. Fossel, A. Liang, M (1992). 'Comparative measurement sensitivity of short and longer health status instruments.' *Medical Care,* 30(10): 917–925.

Kazis, L. Callahan, L. Meenan, R. Pincus, T (1990). 'Health status reports in the care of patients with rheumatoid arthritis.' *Journal of Clinical Epidemiology,* 43(11): 1243–1253.

Kerr, E. Fleming, B (2007). 'Making performance indicators work: experiences of US Veterans Health Administration.' *British Medical Journal,* 335(7627): 971–973.

Lipscomb, J. Gotay, C. Snyder, C (2007). 'Patient-reported outcomes in cancer: a review of recent research and policy initiatives.' *CA: A Cancer Journal for Clinicians,* 57(5): 278–300.

London School of Hygiene and Tropical Medicine (2007). *Patient reported outcome measures (PROMs) in elective surgery*. Report to the

Department of Health. London: London School of Hygiene and Tropical Medicine (http://www.lshtm.ac.uk/hsru/research/PROMs-Report-12-Dec-07.pdf).

Marshall, S. Haywood, K. Fitzpatrick, R (2006). 'Impact of patient-reported outcome measures on routine practice: a structured review.' *Journal of Evaluation and Clinical Practice*, 12(*5*): 559–568.

McHorney, C. Bricker, D (2002). 'A qualitative study of patients' and physicians' views about practice-based functional health assessment.' *Medical Care*, 40(11): 1113–1125.

Moinpour, C. Denicoff, A. Bruner, D. Kornblith, A. Land, S. O Mara, A. Trimble, E (2007). 'Funding patient-reported outcomes in cancer clinical trials.' *Journal of Clinical Oncology*, 25(32): 5100–5115.

Moore, F. Wolfson, C. Alexandrov, L. Lapierre, Y (2004). 'Do general and multiple sclerosis specific quality of life instruments differ?' *Canadian Journal of Neurological Sciences*, 31(1): 64–71.

Moran, L. Guyatt, G. Norman, G (2001). 'Establishing the minimal number of items for a responsive, valid, health-related quality of life instrument.' *Journal of Clinical Epidemiology,* 54(6): 571–579.

Murawski, M. Miederhoff, P (1998). 'On the generalizability of statistical expressions of health-related quality of life instrument responsiveness: a data synthesis.' *Quality of Life Research*, 7(1): 11–22.

Murray, DW. Fitzpatrick, R. Rogers, K. Pandit, H. Beard, DJ. Carr, AJ. Dawson, J (2007). 'The use of the Oxford hip and knee scores.' *Journal of Bone & Joint Surgery*, 89(8): 1010–1014.

Nelson, E. Landgraf, J. Hays, R. Wasson, J. Kirk, J (1990). 'The functional status of patients. How can it be measured in physicians' offices?' *Medical Care,* 28(12): 1111–1126.

Nilsson, E. Wenemark, M. Bendtsen, P. Kristenson, M (2007). 'Respondent satisfaction regarding SF-36 and EQ-5D, and patients' perspectives concerning health outcome assessment within routine health care.' *Quality of Life Research*, 16(10): 1647–1654.

Norquist, J. Fitzpatrick, R. Dawson, J. Jenkinson, C (2004). 'Comparing alternative Rasch-based methods vs raw scores in measuring change in health.' *Medical Care*, 42(Suppl. 1): 125–136.

Pincus, T. Wolfe, F (2005). 'Patient questionnaires for clinical research and improved standard patient care: is it better to have 80% of the information in 100% of patients or 100% of the information in 5% of patients?' *Journal of Rheumatology,* 32(4): 575–577.

Ruta, D. Garratt, A. Leng, M. Russell, I. Macdonald, L (1994). 'A new approach to the measurement of quality of life: the patient-generated index.' *Medical Care,* 32(11): 1109–1126.

Salek, S. Roberts, A. Finlay, A (2007). 'The practical reality of using a patient-reported outcome measure in a routine dermatology clinic.' *Dermatology,* 215(4): 315–319.

Sanders, C. Egger, M. Donovan, J. Tallon, D. Frankel, S (1998). 'Reporting on quality of life in randomised controlled trials: bibliographic study.' *British Medical Journal,* 317(7167): 1191–1194.

Siegrist, J. Wahrendorf, M. von dem Knesebeck, O. Jürges, H. Börsch-Supan, A (2007). 'Quality of work, well-being, and intended early retirement of older employees: baseline results from the SHARE study.' *European Journal of Public Health*, 17(1): 62–68.

Stucki, G. Daltroy, L. Katz, J. Johannesson. M. Liang, M (1996). 'Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts.' *Journal of Clinical Epidemiology*, 49(7): 711–717.

Taylor, K. Macdonald, K. Bezjak, A. Ng, P. DePetrillo, A (1996). 'Physicians' perspectives on quality of life: an exploratory study of oncologists.' *Quality of Life Research,* 5(1): 5–14.

Terwee, C. Bot, S. de Boer, M. van der Windt, D. Knol, D. Dekker, J. Bouter, L. de Vet, H (2007). 'Quality criteria were proposed for measurement properties of health status questionnaires.' *Journal of Clinical Epidemiology*, 60(1): 34–42.

Torrance, GW. Keresteci, MA. Casey, RW. Rosner, AJ. Ryan, N. Breton, MC (2004). 'Development and initial validation of a new preference-based disease-specific health-related quality of life instrument for erectile function.' *Quality of Life Research*, 13(2): 349–359.

Velikova, G. Booth, L. Smith, A. Brown, P. Lynch, P. Brown, J. Selby, P (2004). 'Measuring quality of life in routine oncology practice improves communication and patient well-being: a randomized controlled trial.' *Journal of Clinical Oncology*, 22(4): 714–724.

Wagner, AK. Ehrenberg, BL. Tran, TA. Bungay, KM. Cynn, DJ. Rogers, WH (1997). 'Patient-based health status measurement in clinical practice: a study of its impact on epilepsy patients' care.' *Quality of Life Research*, 6(4): 329–341.

Walsh, TL. Hanscom, B. Lurie, JD. Weinstein, JN (2003). 'Is a condition-specific instrument for patients with low back pain/leg symptoms really necessary? The responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36.' *Spine*, 28(19): 607–615.

Ware, J. Sherbourne, C (1992). 'The MOS 36-item short-form health survey (SF-36).1. Conceptual framework and item selection.' *Medical Care,* 30(6): 473–483.

Wiebe, S. Guyatt, G. Weaver, B. Matijevic, S. Sidwell, C (2003). 'Comparative responsiveness of generic and specific quality of life instruments.' *Journal of Clinical Epidemiology*, 56(1): 52–60.

2.3 *Measuring clinical quality and appropriateness*

ELIZABETH A. MCGLYNN

Introduction

The purpose of this chapter is to review the state of the art in developing clinical process measures and to describe some of the schemes that are using these measures for health system improvement. A high-level summary of the major steps involved in constructing good clinical process measures is provided to enable policy-makers to appreciate some of the complexities involved. There is not enough detail for novices to be able to develop measures from this source alone, but interested readers will be pointed towards examples of best practice.

The section on current schemes that employ clinical process measures includes a greater number of examples from the United States. This reflects the fact that clinical process measurement has been undertaken systematically in the United States for a longer period. Much activity is currently underway in several countries but the measures being used are not readily accessible. Some of these schemes may therefore be under-represented in this chapter.

The chapter concludes with some thoughts on the best uses of process measures, particularly in comparison to outcomes measures. In general, both play an important role in stimulating quality improvement at different levels in the health system and neither type of measure alone is sufficient for all applications. Some directions for future research in this area are also proposed.

State-of-the-art development of clinical process measures

Developers generally pass through five steps to create state-of-the-art measures: (i) selecting topics; (ii) reviewing clinical evidence; (iii) identifying clinical process indicators; (iv) constructing process measures; and (v) creating scoring methods. The importance of each step is discussed below, together with what constitutes best practice.

Selecting topics

Process measurement occurs within a context and the selection of the topics for measurement is a critical step in defining this. The availability and use of performance measures will result in other resources being directed at the measured areas ('what gets measured, gets done') and therefore topic selection should be undertaken systematically. This is particularly important if measures are being developed across multiple clinical areas or for a specific population.

Topics for clinical process measures are generally defined by conditions (e.g. hypertension, upper respiratory infection) although these may be identified for different age groups (children, older people), settings (ambulatory, hospital, nursing home), or events (discharge from hospital, end of life). Ideally, topics are selected because they represent critical dimensions of a strategic plan for improving the health outcomes for a particular group.

The first consideration is to select the outcomes that are of greatest interest – mortality, morbidity, functioning and well-being are the most common outcomes used to identify clinical areas. The availability of systematic data on these outcomes across the group of interest will facilitate topic selection. Mortality data are the most likely to be available (through national data systems) followed by morbidity. Systematic, national (or system-level) information on functioning and well-being are much less likely to be available. Data collection may be best informed by a review of published studies or through a group process that obtains input from experts or community leaders.

A second consideration is the condition's relative impact on the population of interest. In general, priority is given to conditions that are highly prevalent (e.g. the top ten causes of death) or have a substantial impact on health (e.g. those with the condition have a very high probability of dying). For example, heart disease and cancer are the leading causes of death in the United States and so would be high priorities to support plans for reducing premature mortality. Severe depression is one of the leading causes of functional limitations worldwide and would likely be included in a strategy to improve functioning.

A third consideration is whether outcomes are likely to be affected by actions taken in the health-care system. A number of potential actions to improve population health do not operate through the

health-care system (e.g. ensuring adequate sanitation, safe food, clean environments) and some areas do not have health services that are effective in changing an outcome. Neither of these areas is fruitful for developing clinical process measures.

There are a number of examples of systematic selection of topics for quality improvement or measurement. For example, the Institute of Medicine (2003) identified twenty priority areas for quality measurement representing clinical areas across the age spectrum. The Danish National Indicator Project selected clinical areas representing the greatest use of resources in hospitals (Mainz et al. 2004). The Assessing Care of Vulnerable Elders (ACOVE) project selected twenty-six conditions representing clinical problems of the elderly population using a group judgment process (Wenger et al. 2007).

Reviewing clinical evidence

Once a topic (or topics) has been selected, the next step is to review what is known about effective interventions. The starting point is to construct the questions that will be answered by the literature search. For example, if heart disease was selected as a topic the questions might include the following.

- What interventions have been shown to be effective in preventing heart disease (primary and secondary prevention)? What interventions have been shown to be ineffective in preventing heart disease?
- Is there evidence that early identification of heart disease through general population screening reduces premature mortality or morbidity, or leads to higher functioning?
- What methods are effective in accurately diagnosing the presence (or absence) of heart disease? What methods are not effective or are unnecessary aspects of the diagnostic process?
- What interventions have been shown to be effective in treating established heart disease? What interventions have been shown to be ineffective in treating heart disease?
- What interventions have been shown to be effective in helping people return to higher levels of functioning following a heart attack? What interventions have been shown to be ineffective?

- What interventions have been shown to be effective for the ongoing management of persons with established heart disease? What interventions have been shown to be ineffective?

In this example, separate questions are posed for primary and secondary prevention, screening, diagnosis, treatment, rehabilitation and ongoing maintenance. This is appropriate for developing measures across the continuum of care but measures focused on a single aspect of the continuum would require only questions related to that area. Positive (what is effective) and negative (what is not effective) questions are asked to illustrate how evidence for measures of underuse (failure to use effective interventions) or overuse (use of interventions known to be ineffective) might be developed.

A formal strategy for identifying relevant articles is developed once the questions have been agreed upon. Several components are involved and the choices within each will depend on the time and resources available; degree to which an exhaustive search is necessary to meet the goals; and the likelihood of reaching a different conclusion by broadening the search strategy. This must include consideration of the type of studies that will be included (e.g. only randomized trials or a broad range of study designs); whether particular outcomes have been measured (e.g. include only studies that examine the impact on premature mortality or on functioning); the characteristics of participants (e.g. development of measures for the elderly might require only studies on this population); and what specific interventions are included (e.g. only those that can be provided in ambulatory settings vs. any setting). In addition, the reviewer must consider what databases to search; how far back to look; whether to supplement electronic searches with other information (e.g. literature cited in articles, hand search of specific journals); or whether to include information that has not been published in peer reviewed journals (e.g. private reports, data from unpublished studies).

Generally, the articles that will be included in a review are determined via three steps. First, a list of article titles is obtained by the application of search terms and other strategies. This list is screened to identify those that are relevant for the particular question and to exclude those that are not. Second, these selected titles undergo a more formal screen of abstracts to further determine which of these should be included. This step can be used to apply some of the selection

criteria (e.g. type of study, population, outcomes). Third, a full review is conducted on the articles selected during the abstract review and relevant information is collated. Some articles may be excluded at this step if greater detail available in the full article indicates that they do not meet the inclusion criteria (or do meet the exclusion criteria). These review results are generally summarized in an evidence table.

Clinical practice guidelines are another source of evidence for constructing process measures. Evidence-based guidelines will incorporate conclusions from the scientific research literature about preferred approaches to prevention, screening, diagnosis, treatment, rehabilitation and monitoring. Even evidence-based guidelines will include some guidance that reflects professional consensus rather than scientific studies. Well-documented guidelines should enable the reviewer (or user) to identify easily the foundation for each recommendation (Shiffman et al. 2003). The Agency for Healthcare Research and Quality (AHRQ) maintains the National Guideline Clearinghouse (www.guideline.gov). This holds guidelines from a variety of sources and currently has 2083 individual summaries from eight countries; European medical societies and WHO. A search for myocardial infarction identified 252 related guidelines, ranging from those providing guidance on the use of a single technology (e.g. electrocardiographic monitoring in the hospital setting) to the management of a diagnosis (e.g. ischaemic heart disease). It is not unusual to find some disagreement between guidelines developed by different groups and these may be worth noting because of their potential impact on the development of process measures.

Two important principles should be kept in mind when developing the evidence base for process measures. First, it is important to document the strategy used to retrieve articles because it allows others to replicate the approach. Second, it is important to consider how the review of evidence might be biased. For example, the search for unpublished literature is designed to deal with publication bias – studies that report positive findings will be published more often than those that report negative or no findings.

The approach described here is consistent with the practices for identifying articles used by the Cochrane Collaboration (http://www.cochrane.org/reviews/revstruc.htm), the AHRQ Evidence-based Practice Centers (http://effectivehealthcare.ahrq.gov/) and the National Institute for Health and Clinical Excellence (NICE) (http://www.nice.

org.uk/guidance/index.jsp). The process is similar to that used in creating evidence-based guidelines.

Identifying clinical process indicators

Process indicators are descriptive statements about the aspect of care that is being evaluated and the type of patient that should receive the indicated care. Most clinical process indicators are written in a general style, such as:

- Persons with diabetes mellitus should have their blood sugar measured at least once each year.

The style introduced by the ACOVE project makes the eligibility and expected process statements more explicit by using if/then statements in which the 'if' describes the eligible population and the 'then' describes the expected care process (Wenger et al. 2007). For example:

- *If* a vulnerable elder has diabetes mellitus, *then* glycated haemoglobin (HbA1c) should be measured annually.

Clinical practice guidelines are 'systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances' (Institute of Medicine 1990). Clinical process indicators have a different purpose as they are designed to guide the evaluation of health service delivery. As a result, they have some key distinguishing features:

- selective rather than comprehensive;
- usually focus on areas for which a link to outcomes has been established in the scientific literature;
- inclusion and exclusion criteria are explicit rather than left to clinical judgment;
- intended to apply to the average patient seeing the average physician;
- applied retrospectively to a population of patients (guidelines are used prospectively in the management of a single patient).

Process indicators should be selected in a way that maintains a link to the evidence that supports the underlying scientific rationale. The RAND/UCLA Appropriateness Method established an approach to selecting indicators that combines a review of published evidence with a formal expert panel process (Brook 1994). This method is

reliable (Shekelle et al. 1998) and has been shown to have content, construct and predictive validity in other applications (Hemingway et al. 2001; Kravitz et al. 1998; Selby et al. 1998; Shekelle et al. 1998a).

In this approach, the development staff produces a set of draft quality indicators based on a review of the literature and guidelines (as described above) and measurement expertise. An expert panel is recruited based on nominations from appropriate specialty societies. The panels generally have nine doctors and include multiple specialties (e.g. primary care and specialty care doctors, proceduralists and non-proceduralists) and are diverse with respect to geography, gender, practice setting and other factors relevant to the purpose of the quality indicator set.

The draft process indicators and the literature review described above are referred to an expert panel (usually of nine members) that votes on which indicators should be included. Each panel member rates each indicator privately on a scale from one (i.e. indicator is not a valid measure of quality) to nine (i.e. indicator is a very valid measure of quality). The development staff summarizes results from the initial round of ratings for each indicator to produce the median score on validity (central tendency) and the mean absolute deviation from the median (spread) and to show whether the indicator ratings demonstrate substantial agreement or disagreement. Panellists assemble to discuss the indicators in a face-to-face meeting that allows them all to benefit from the perspectives of those with different views. Discussion usually focuses on the indicators for which there was substantial disagreement in the first round of ratings (for a nine-member panel defined as three or more ratings ≤ 3 and three or more ratings ≥ 7).

There are two common reasons for disagreement. First, if the indicator language is unclear the panellists may interpret the intent differently. In this case staff can rewrite the indicator or clarify definitions for key terms so that all panellists consider the same group of patients in their ratings. Second, the indicator may address a clinical process for which no strong evidence or consensus exists. In this case the indicator is likely to be rejected because reasonable people could disagree and there is no strong case for choosing one process over another. Panellists vote again after the group discussion and these results determine which indicators will be included. The standard for the RAND/UCLA method is to include indicators with a median validity score of seven or more that are rated without disagreement.

This method can be used to create appropriateness of care indicators as well as process quality indicators. The panel process is described in detail in the volumes on the RAND Quality Assessment (QA) Tools measures (Asch et al. 2000; Kerr et al. 2000; Kerr et al. 2000a; McGlynn et al. 2000; McGlynn et al. 2000a). Although it is common for countries to conduct their own indicator selection processes, they frequently refer to indicators that have been developed elsewhere. Many indicators transfer well from one country to another because the scientific basis is often common internationally (Steel et al. 2004). However, transferability may be limited by the organization of the delivery system in a country, as was noted in the development of the German indicators for the quality of acute stroke care (Heuschmann et al. 2006).

Constructing process measures

Ideally, the data source is decided prior to the development of process indicators as the type of data available will determine the types of indicators that can be constructed. When this does not happen some indicators will likely be dropped during measure development because it will not be feasible to collect the necessary data from the intended source.

Data sources

There are three major sources of data for measuring process quality: (i) medical records (electronic or paper); (ii) billing data; and (iii) surveys (patient or doctor). Each of these has strengths and weaknesses which limit the types of indicators that can be evaluated and the validity of results. Some of the main considerations are highlighted in the following paragraphs but there is not sufficient space for full descriptions of all.

Medical records contain the greatest amount of clinical information and allow the construction of measures that are clinically detailed with respect to defining eligibility, exclusions and scoring criteria. Collecting data from paper-based medical records is labour intensive and this may limit their utility for routine assessments. Paper-based medical records also lack standardized nomenclature which means that data collectors need to be carefully trained and supervised to

ensure that results are reliable and valid across providers. Electronic medical records may offer greater ease of access but many such systems face the same limitations as paper-based records (lack of standard nomenclature, need to abstract key pieces of information manually). Developers (and purchasers) of electronic medical records systems face difficult tradeoffs between ease of implementation for users and the utility of the information produced for secondary uses. To date ease of use by clinicians (which may be necessary to stimulate adoption of the technology) has been prioritized.

Billing data have the advantages of being available electronically and constructed using standardized coding schemes but they lack clinical detail. In most cases a bill indicates that an encounter took place but contains no information on its content, apart from separately billed interventions (e.g. laboratory tests, immunizations, other procedures). Also, there is usually no information about the clinical profile of the patient (e.g. severity and extent of disease, co-morbid conditions, behavioural risk factors). Thus, billing data are most useful for quality indicators that require little clinical detail to identify the eligible population (e.g. presence of disease is sufficient and exclusions are rare) and to determine whether the process occurred (e.g. whether a laboratory test was ordered rather than whether counselling about a health-related behaviour occurred). Most billing data do not include the results of tests ordered (e.g. HbA1c level; LDL cholesterol level; imaging) although such information is increasingly becoming available electronically and integrated into data warehouses.

Patient survey data are useful when the patient is a reliable reporter about the eligibility conditions (e.g. presence of disease, age, health risk behaviour, symptoms) and whether or not a process occurred (e.g. various screening tests, advice from a doctor). Patients have more difficulty reporting specific test values although they may be aware of whether intermediate outcome measures for chronic diseases (e.g. blood pressure, blood sugar, cholesterol) are high, low or normal. Patient surveys can be difficult to collect on a representative sample because people are unwilling to participate or may be hard to reach.

Surveys of doctors are useful when evaluating knowledge about particular care processes or using scenarios to test what the doctor might do. Doctors are less likely than patients to respond to surveys. Studies have shown that knowledge does not necessarily translate into action so knowledge-based surveys may not be indicative of actual

performance. Scenario-based studies are more reliable but there is a limit to the number of scenarios that can be tested in a single survey.

Development of measures

Detailed specifications must be developed to enable reliable assessment of the frequency with which a clinical process is delivered. The specifications should define unambiguously the criteria for identifying patients who are eligible for a clinical process indicator and for determining whether eligible patients received the indicated care. The specifications will take different forms depending on the data source. To illustrate the approach to developing specifications, consider the following indicator:

- Persons with diabetes mellitus should have their blood sugar measured at least once per year.

The first step is to develop specifications to identify those who have diabetes mellitus. It is common to consider first whether the eligible population needs to be restricted in any way. This is illustrated in Table 2.3.1 below.

These questions illustrate a major point in constructing process quality measures. In general, such measures are designed with a tendency to specificity rather than sensitivity: appropriate for the population identified as eligible for a measure (with only rare exceptions) to receive the care process. The difference between these considerations and a clinical guideline is that guidelines can allow for clinical judgment – the doctor is responsible for determining the tradeoffs based on knowledge of the patient's full spectrum of health concerns.

To ensure that data are collected reliably for a process measure, the data collector's judgment must be largely removed. It is rarely possible to include all possible clinical exceptions to eligibility. It may not be necessary to include an exception that is 'rare or random' but those that are 'common or biased' should likely be included. This requires consideration of the application of the measure by asking whether a particular clinical exception occurs less than, for example, 1% of the time in the population of interest and whether the exceptions would be expected to be distributed randomly (without any discernable pattern) across the entities likely to be evaluated. If that test is met, it may not be worthwhile to include exceptions for those considerations.

Table 2.3.1 ***Assessing the eligible population for clinical process indicators***

Consideration	Issue for measurement
Does the measure apply to patients at all ages or should lower and/or upper age limits be established?	There may be ages (e.g. children, older adults) where the clinical judgment is more critical than the standard reflected in the indicator.
Does the measure apply to both type 1 and type 2 diabetes?	Subgroups within a diagnosis may be excluded. In this case, it is often difficult to distinguish between type 1 and type 2 in various data sources and routine measurement of blood sugar is the standard of care for both.
Does the measure apply to women with gestational diabetes?	The nature of the diagnosis and the routine management indicators are different for this subgroup.
Does the measure apply only to patients with a confirmed or established diagnosis of diabetes or can other factors (e.g. high HbA1c value) be used to identify the eligible population?	When assessing potential underuse it is sometimes appropriate to include persons who have signs of disease but diagnosis has not been recognized in the medical record. The conservative approach requires a confirmed diagnosis.
Should persons with a new diagnosis be included or does the patient have to have had the diagnosis for some period of time?	Process measures often distinguish between new diagnoses (where measures related to the quality of the diagnostic process are appropriate) and prevalent or existing diagnoses (where routine management measures are appropriate). This indicator is intended to apply to those with an established diagnosis.
Should there be exclusions for co-morbid conditions?	In some cases management of a co-morbid condition (cancer, AIDS) will take priority over routine care for the condition under consideration for process measures.

Table 2.3.1 ***(cont)***

Consideration	Issue for measurement
Should there be exclusions for health status (e.g. end of life)?	Similar to the above consideration, some routine management of chronic conditions will be inappropriate at the end of life or in the face of other health status concerns.

The inclusion criteria are the next consideration and some of the questions that typically determine these are shown in Table 2.3.2.

Together, the exclusion and inclusion criteria form the basis for identifying the eligible population but the way in which these considerations are operationalized varies with the data source. Medical records require instructions related to notation; billing data require instructions that include the common codes used (e.g. ICD-9 or 10, CPT-4); and patient surveys need a set of questions that will elicit information about the inclusion and exclusion criteria. Most process measures are constructed by putting together a sequence of events and determining whether these occurred within an acceptable time frame. For this reason, it is generally better to collect the date associated with an event (e.g. visit for diabetes) rather than a dichotomous answer (yes/no) to a question about whether criteria are met. This allows maximum flexibility in assessing whether an indicated process has been met.

Finally, the criteria for determining whether or not the indicated care has been delivered are specified. The questions in Table 2.3.3 illustrate how this might be done.

The specifications must include instructions about the type of documentation or the names of laboratory tests that meet the conditions. Specific codes must be listed if billing data are being used.

Creating scoring methods

The last major development step is to create the scoring instructions. For process measures, the basic approach for an individual indicator is to count the number of times that a patient in the population of interest is eligible for an indicator and then the number of times the

Table 2.3.2 ***Determining the inclusion criteria for clinical process indicators***

Consideration	Issue for measurement
What evidence is sufficient to determine that the patient has diabetes?	Options for a chronic disease include: (i) visit where the reason for visit is the diagnosis; (ii) medication orders consistent with the diagnosis (insulin, oral hypoglycaemics); (iii) mention of diabetes as a co-morbid factor in a visit for another reason.
Will the measure be limited to those with evidence of the disease in the year in which the measure is constructed?	When looking for evidence of underuse, and when the diagnosis is not likely to resolve, evidence of disease in a time period prior to the one in which the care process is being evaluated is acceptable. The look-back period may be limited in order to improve data collection efficiency.

indicated process was delivered to those who are eligible. Table 2.3.4 illustrates this process for a simple indicator with five patients in the population of interest.

In this example, four of the five patients are eligible for the indicator and two passed for a score of two out of four, or 50%.

Some process indicators require multiple events. For example, if the example indicator requires two blood sugar tests per year it needs to be decided whether the scoring method is 'all or nothing'. This means that the indicator is not passed if a patient receives fewer than two tests in a year. Alternatively, partial credit can be granted by counting the proportion of required tests received. This can be seen as giving each patient two eligibilities (one for each test that should be received) and counting the number of times the process was received. The scores for an individual patient would be 0%, 50% or 100%.

Increasingly, process indicators are being combined to create composite scores. For example, a diabetes composite score could be compiled from multiple process indicators related to routine management of diabetes. Similarly, composites can be created across conditions (for example, all chronic disease care in a population). Composites are constructed in three common ways.

Table 2.3.3 ***Specifying the criteria for determining whether indicated care has been delivered***

Consideration	Issue for measurement
What type of blood sugar test is sufficient to meet the conditions for the indicator?	This has generally been limited to an HbA1c test, but multiple tests might be allowed to meet a criterion for other indicators. Possible question for this indicator is whether home monitoring tests are an acceptable alternative – they would not be accepted as they do not meet the intent of this indicator.
Is there evidence that an HbA1c test was ordered or that laboratory results are available?	Tests (and medications) have two signals – whether the test was ordered and whether it was completed. Accounting for orders gives the doctor the benefit of the doubt, particularly in systems where the patient goes elsewhere for the test. Alternatively, orders may not be recorded in some records but laboratory reports show that a test was done. Standard practice at RAND is to take account of both orders and test results.
Is there evidence that the patient refused the test?	Look for documentation that the patient refused a recommended procedure (only possible in medical record-based data collection or surveys) and allow refusals to count toward passing an indicator. Refusals could also be used to exclude a patient from an indicator.
Does the sequence of events matter?	Some instances may require evidence that a diagnosis occurred on or before the date of the indicated process (blood sugar test). Here, the sequence is not important because prevalent cases of diabetes are sought. Those with a new diagnosis have been excluded. However, this type of consideration illustrates why it is useful to have the dates on which events occur.

Table 2.3.4 ***Sample scoring table for a simple performance indicator***

Patient	Eligible?	Received process?
1234	Yes	Yes
5678	No	NA
9101	Yes	No
1112	Yes	No
1314	Yes	Yes

1. Opportunity score counts all instances in which a patient is eligible for an indicator in the denominator and all instances in which the indicated care was delivered in the numerator. The implicit weight in this case is the prevalence of eligibility for different indicators – more common care processes account for a greater portion of the total score and patients who are eligible for more indicators contribute more to the total score.
2. Average of averages approach creates a score for each patient and then averages the patients' scores. In this case, each patient counts equally toward the total score.
3. All or nothing approach counts the proportion of patients who receive all the care for which they are eligible. Each patient counts equally although patients eligible for a larger number of indicators may be less likely to get all indicated care.

Weights can be added within each of these general approaches in order to reflect the different levels of clinical importance attached to certain indicators.

Risk adjustment is used less commonly for process measures. The rationale is that most of the risk adjustment occurs in constructing the conditions of eligibility. If the process measures are being used to compare the performance of different entities, this might include consideration of whether one entity has a greater number of patients or eligibility events associated with indicators that have low empirical scores (i.e. appear to be harder to pass). At RAND, adjustments to scores have been constructed to account for this.

Process measurement schemes in operation

Process measures can be used in a variety of ways to improve quality, for example as part of accreditation of facilities or providers; in public reporting; as part of the structure of benefit designs; and in payment incentive programmes. In this section, some of the current uses of process measures outside of the research setting are described. This is not exhaustive but is intended to illustrate some of the ways in which clinical process and appropriateness measures can be used to promote quality.

Accreditation

Accreditation is the recognition by an independent body that an organization meets an acceptable standard. Traditionally, accrediting organizations have set standards related to the way in which an organization functions (e.g. whether specific procedures are in place, certain committees exist and meet regularly, safety codes are met) and assessed compliance with these standards through on-site visits. It is less common for accrediting bodies to use process measures to assess actual performance.

In the United States, the National Committee for Quality Assurance (NCQA) uses about twenty measures of process quality as part of its accreditation programme for managed care and preferred provider organizations. A description of NCQA's accreditation programme is available on its web site (http://www.ncqa.org/). The process measures selected for accreditation are drawn from the Healthcare Effectiveness Data and Information Set – HEDIS (Lacourciere 2007). They meet the best practice for measure development described above. The process quality and patient experience measures account for about 40% of the total accreditation score. Managed care organizations participate in accreditation voluntarily but about 90% of such organizations in the United States seek NCQA accreditation.

At the time of writing, no European countries were identified that had incorporated clinical process measures into any voluntary accreditation schemes. To the extent that accreditation is used in Europe, the performance measures included are more likely to relate to the volume of procedures performed or the waiting times to access a procedure. Sometimes volume is used as a proxy for quality but it is not consid-

ered a clinical process measure. In this context waiting times also do not constitute a clinical process measure.

Public reporting

The results of clinical process and appropriateness assessments have been reported at various levels in the health-care system. National reporting is perhaps the most common and in recent years there has been an interest in common measures that allow for cross-national comparisons. Results can also be reported anonymously or by the name of the provider (health plan, hospital, nursing home, medical group physician).

The AHRQ has produced an annual report on health-care quality since 2003. The 2008 report is available from the web site (http://www.ahrq.gov/qual/qrdr08.htm#toc). A variety of data sources are used to construct the measures which report on the following clinical areas: cancer; diabetes; end stage renal disease; heart disease; HIV and AIDs; maternal and child health; mental health and substance abuse; and respiratory diseases. Process indicators constitute the largest portion of the indicators.

The Organisation for Economic Co-operation and Development (OECD) is conducting a project to collect national-level information on process quality suitable for cross-national comparisons. The project started in 2001 and involves twenty-three countries. The initial report contained seventeen indicators, primarily outcomes measures but including some process quality indicators for cancer screening (breast, cervical) and vaccinations (childhood, adult influenza) (Mattke et al. 2006). An indicator for retinal screening among persons with diabetes has been added subsequently (OECD 2007).

There are a number of examples of public reporting for managed care organizations, hospitals and nursing homes in the United States. Some organizations are also working to develop public reports of performance at the medical group practice and individual doctor level. Since 1999, NCQA has released public reports using clinical process measures. A subset of the information collected by NCQA is available on the web site (http://hprc.ncqa.org/tabid/836/Default.aspx) and more detailed information can be purchased. The web site provides a high level summary (one to four stars) of performance in a category

(e.g. chronic disease category = living with illness) and scores for a subset of eleven individual measures for asthma, diabetes, heart disease and mental health are available. These results are shown for each health plan along with a comparison to the score for each measure for the top 10% of plans nationally and the top 25% and 50% of plans regionally.

Public reports on hospital performance in the United States are available from the Centers for Medicare and Medicaid Services (http://www.hospitalcompare.hhs.gov/). Bar graphs show the results for three clinical areas (heart attack, heart failure, pneumonia) and for surgical care (prevention of infections). Results are displayed for hospitals selected by the user and are compared to the United States' average, the average for the state in which the hospital is located and the top 10% of hospitals nationwide. The information is also available as a table that includes the number of patients who were eligible for the measure. The Joint Commission provides reports on the same measures in a different format – symbols provide a high-level summary of performance in the category and detailed information is provided on each process measure within the category available. There are comparisons with the top 10% and average scores both nationally and for the state in which the hospital is located (http://www.qualitycheck.org/).

The Netherlands Health Care Inspectorate (IGZ) has developed a set of hospital performance indicators that include a combination of structure, process and outcomes measures (Dutch Institute for Healthcare Improvement 2004). The process measures are based on national guidelines. The Danish National Indicator Project focuses on hospital-delivered care in eight clinical areas: stroke, hip fracture, upper gastrointestinal bleeding, lung cancer, schizophrenia, heart failure, diabetes and chronic obstructive lung disease (Bartels et al. 2007). Participation in reporting is mandatory for hospitals and the results are reported using both opportunity and all-or-nothing scoring methods.

Process measures are not used as commonly in public reports on nursing homes in the United States. These include only two process measures – on influenza and pneumococcal vaccinations (http://www.medicare.gov/NHCompare). Public reports of performance on clinical process measures have been available at the medical group and clinic level in Minnesota for the past four years through a private nonprofit group. MN Community Measurement was founded by the Minnesota Medical Association and seven of the nonprofit health plans operating in the state (http://www.mnhealthscores.org/Report/). The reports

include measures of care processes and outcomes in nine clinical areas: asthma, cancer screening, childhood immunizations, chlamydia screening, diabetes, pharyngitis (sore throat), upper respiratory infection, vascular care and coronary artery disease care. Reports for the optimal care measurement areas (diabetes, cancer screening, vascular and coronary artery disease) use the all-or-nothing scoring method and are dominated by outcomes measures.

In the United Kingdom, the Quality and Outcomes Framework (QOF) uses process measures to assess the performance of general practices. The clinical domain currently includes eighty indicators across nineteen clinical areas. The results are available on multiple web sites providing overall statistics for the nation (proportion of practices achieving 100% performance, average performance levels) and an online database that allows users access to detailed information about specific practices. The online database (http://www.qof.ic.nhs.uk/) has a number of display options including comparisons between a selected practice and the averages for the local primary care trust (PCT) and England, respectively.

Benefit design

The use of process measures for benefit design is a relatively new phenomenon in the United States. Essentially, process measures are used to assess the relative performance of hospitals, medical groups or physicians. Patients pay copayments based on relative rankings – lower copayments are due if patients see providers with relatively better performance. The purpose is to provide patients with a financial incentive to seek care from better quality providers. These schemes are used by both private insurance companies (e.g. UnitedHealthcare, Aetna) and in government run programmes (e.g. the General Insurance Commission for the state of Massachusetts). In these schemes, process quality measures are generally combined with measures related to the cost of care and the most favourable copayments are assigned to providers who deliver high quality care at low relative cost.

Payment incentives

Process measures have also been used as the basis for payment incentives for providers. These schemes are commonly referred to as pay-

for-performance programmes and have been implemented at hospital and medical group level in the United States and at practice level in the United Kingdom.

About twenty-three hospital pay-for-performance programmes currently operate in the United States. Most draw on the process measures that the Centers for Medicare and Medicaid Services require for reporting (heart attack, heart failure, pneumonia, surgical infection prevention). Typically, composite scores are constructed at the condition level and hospitals are eligible for bonuses (lump sum or percentage) based on the level of achievement. For example, the Premier Hospital Quality Incentive Demonstration paid a 2% bonus to hospitals in the top decile and a 1% bonus to hospitals in the second decile (Lindenauer 2007).

The Integrated Healthcare Association in California has one of the longest running pay-for-performance schemes in the United States. The programme is designed to incentivize medical groups to improve quality. About half of the payment incentive is based on quality measures and eight of the ten measures used in 2005/2006 were clinical process measures (Integrated Healthcare Association 2006).

The pay-for-performance scheme for general practices in the United Kingdom has the most extensive use of process quality measures to date. At the outset of the programme, the government increased the amount of funding for general practices by more than £ 1 billion, an approximately 20% increase in general practice budgets (Roland 2004). Incentives are based on a complex formula that includes minimum and maximum thresholds of performance and a number of points allocated for each indicator. Practice size and the prevalence of different chronic disease are also included in the calculations. At the beginning of the scheme each point was worth £120.

Best uses of process measurement

Most schemes to monitor quality include a combination of different types of measures – structure, process and outcomes. This is reasonable because no single approach to quality measurement addresses all issues. Measures should be selected after consideration of the intended use of the results as this may inform the type of measure preferred.

Process measures have four main advantages. First, care processes occur more frequently thereby enabling deficits in care to be identified

more rapidly. Many quality measurement schemes encounter too few cases to be able to draw robust conclusions; a problem that tends to be more pronounced with outcome measures. Second, process measures describe the care delivery expectations and thus define what needs to be done to achieve optimal care delivery. When monitoring outcomes, the reasons for poor results are not always clear and it may be necessary to collect process measures to identify the steps that must be taken to improve these outcomes. Third, process measures generally do not require risk adjustment beyond the specifications associated with identifying eligible patients. This increases the potential for greater acceptability of the measures as risk adjustment of outcome measures is challenging (and rarely satisfies those being measured). Fourth, processes reflect the way in which the scientific literature is organized. Most studies involve investigations of the effect of a particular intervention and allow direct links to an evidence base.

So when are these attributes most important? As a general rule, process measures are preferred when quality is being measured for the purpose of holding organizations or individuals accountable for meeting standards. This is particularly true when organizations or individuals are being compared.

Recommendations for developing countries

Increasing research has been conducted to investigate the clinical quality improvement efforts being undertaken in developing countries. Successful efforts that have been documented show that these use similar meta-analyses to those undertaken in developed countries (Leatherman et al., forthcoming). Some conclusions on successful monitoring of quality in settings with limited resources can be drawn from projects that have shown favourable outcomes (Berwick 2004; Ovretveit 2004).

Developing countries face the major barrier of a lack of available resources for quality measurement and monitoring. This makes it more difficult to introduce not only the infrastructure necessary for measurement and monitoring, but also staff training and supervision programmes. Yet, investment in these areas has long-term potential as it will enable gaps in quality to be identified and addressed to produce more efficient allocation of financial resources.

Given the limited information technology available in developing countries it is important to measure only what is necessary to inform

policy and not to waste resources by attempting to measure too much. Furthermore, quality measurement and monitoring should be directed at areas in which quality improvements will have the most impact. These may differ from the clinical areas targeted by developed countries. For example, much of the current literature on quality improvement in developing countries describes efforts in the areas of acute illnesses, child care and maternity care. Such efforts may result in increases in immunization rates or reductions in childhood and maternal mortality which have a larger impact on the mortality and morbidity of developing countries.

Where there is a distinct lack of infrastructure, managers should be encouraged to think innovatively about alternative ways of measuring quality. Berwick (2004) gives the example of a maternal and child health clinic in northern Pakistan that wanted to measure the effect of a project on early intervention in pregnancy. The lead doctor on the project suggested counting the small graves as an outcome measure. This shows a creative way of overcoming the lack of IT infrastructure to address the problem at hand.

Physical infrastructures need to be developed in tandem with training programmes that provide all levels of staff with the skills to carry out a systematic measurement of indicators. Moreover, teamwork should be encouraged amongst those employing interventions at the provider, patient and system levels to ensure that measurement and monitoring is integrated throughout.

Finally, it may be useful to develop different systems to reward the practitioners or facilities that undertake quality monitoring. These can take the form of self-assessment, peer review, certification, accreditation or licensing. Such mechanisms allow recognition of more successful endeavours as well as the identification of areas where quality monitoring efforts are less effective and can be improved.

Directions for future research

Clinical process measures offer an important tool for assessing the current quality of care being delivered by a system or in a country. They are also useful for evaluating whether interventions have improved quality performance. This chapter has described the challenges associated with developing robust process measures and with implement-

ing assessments on a large scale. One promising direction for future research is the development of streamlined approaches to measure the development and translation of measures across systems and countries. This activity may be performed most effectively by a limited number of centres with special expertise in combination with government or nongovernmental organizations that translate measures into routine use. Methods to assess the appropriateness of a measure or set of measures for use in a new country or system could increase the potential to use or adapt measures in new settings.

There is considerable interest in cross-national comparison of quality performance. Much of what is known today is derived from surveys but a number of clinical process measures cannot be assessed adequately in this way. The development of a core set of process measures that could be used across countries with different health systems would increase the ability of countries to learn from one another. This would likely require investment from a group that takes the lead on this activity as well as cooperation from participating countries. Such efforts are underway but have encountered considerable difficulties.

Another critical area for research is to find ways to integrate measurement and clinical practice. Too often, quality measurement activities are separate from the delivery of health services. Quality will not reach its full potential until methods for measuring and delivering care can be integrated.

It would be useful to identify a set of strategies that are effective in improving quality in different settings and countries. Research in this area is fairly rudimentary and requires considerable work to identify the best ways of converting information generated from process assessments into action plans for improvement. With few effective ways of sharing lessons learned across different entities and countries, much time is spent on unnecessary duplication.

Quality measurement in developing countries offers an opportunity for innovative thinking and approaches that could well translate to developed countries. Developing countries may offer fresh perspectives on common quality problems and be less tied to a history of how such problems have been solved. It should be a high priority to find ways to draw upon the lessons learned from these experiences and to make this learning widely available.

Conclusions

The methods for developing clinical process and appropriateness measures are well established and the use of state-of-the-art methods has been demonstrated in multiple countries. There has been a substantial increase in the number of measures available but their use for quality improvement and other applications remains limited. The United States appears to lead the world in the use of clinical process measures in different applications, although the United Kingdom's pay-for-performance scheme is far more comprehensive. It is beyond the purview of this chapter to comment on how effective these measures have been in stimulating quality improvement but examples from several countries show positive trends.

One of the greatest limitations to the rapid uptake of clinical process measures is the inadequate data infrastructure in place to support measurement. Health care lags behind most modern industries in its use of electronic systems for the management of essential processes. Without this type of infrastructure, quality measurement is likely to be relegated to a minor role and is unlikely to realize its full potential. Significant investments will be required to develop the necessary information infrastructure to manage patients effectively in the face of accelerating advances in knowledge as the cognitive processes necessary to process the match between patients' problems and the available solutions exceed human abilities. A by-product of this investment will be the development (if done well) of systems that will also allow quality measurement to accelerate. It will be necessary to take account of the information requirements for clinical process measurement as the functional requirements for future health-care information systems are developed.

References

Asch, S. Kerr, EA. Hamilton, EG. Reifel, JL. McGlynn, EA (eds.) (2000). *Quality of care for oncologic conditions and HIV: a review of the literature and quality indicators*. Santa Monica, CA: RAND Corporation (Publication No. MR-1281-AHRQ).

Bartels, PD. Mainz, J. Hansen, A-M. Ingeman, A. Bunk, A. Nakano, A. Kaersvang L (2007). *Nationwide performance measurement can*

improve the quality of care. Presentation at International Society for Quality in Health Care, 1 October 2007 (http://www.isqua.org/isqua-Pages/Conferences/Boston/slides/BOSTONMonday.html).

Berwick, DM (2004). 'Lessons from developing nations on improving health care.' *British Medical Journal*, 328(7448): 1124–1129.

Brook, RH (1994). The RAND/UCLA Appropriateness Method. In: McCormick, KA. Moore, SR. Siegel, RA (eds.). *Clinical practice guideline development: methodology perspectives*. Rockville, MD: Agency for Health Care Policy and Research, Public Health Service, US Department of Health and Human Services (Publication No. 95–0009).

Dutch Institute for Healthcare Improvement (2004). *Vision of quality 2004*. Utrecht: Dutch Institute for Healthcare Improvement (http://www.cbo.nl/algemeen/visionofquality.pdf/view).

Field, MJ. Lohr, KN (eds.) (1990). *Clinical practice guidelines: directions for a new program*. Washington, DC: Institute of Medicine National Academies Press.

Hemingway, H. Crook, AM. Feder, G. Banerjee, S. Dawson, JR. Magee, P. Philpott, S. Sanders, J. Wood, A. Timmis, AD (2001). 'Underuse of coronary revascularization procedures in patients considered appropriate candidates for revascularization'. *New England Journal of Medicine*, 344(9): 645–654.

Heuschmann, PU. Biegler, MK. Busse, O. Elsner, S. Grau, A. Hasenbein, U. Hermanek, P. Janzen, RW. Kolominsky-Rabas, PL. Kraywinkel, K. Lowitzsch, K. Misselwitz, B. Nabavi, DG. Otten, K. Pientka, L. von Reutern, GM. Ringelstein, EB. Sander, D. Wagner, M. Berger, K (2006). 'Development and implementation of evidence-based indicators for measuring quality of acute stroke care: the Quality Indicator Board of the German Stroke Registers Study Group (ADSR).' *Stroke*, 37(10): 2573–2578.

Institute of Medicine (2003). *Priority areas for national action: transforming health care quality*. Washington, DC: National Academy Press.

Integrated Healthcare Association (2006). *Advancing quality through collaboration: the California pay for performance program*. Oakland, CA: Integrated Healthcare Association.

Kerr, EA. Asch, S. Hamilton, EG. McGlynn, EA (eds.) (2000). *Quality of care for cardiopulmonary conditions: a review of the literature and quality indicators*. Santa Monica, CA: RAND Corporation (Publication No. MR-1282-AHRQ).

Kerr, EA. Asch, S. Hamilton, EG. McGlynn, EA. (eds.) (2000a). *Quality of care for general medical conditions: a review of the literature and quality indicators*. Santa Monica, CA: RAND Corporation (Publication No. MR-1280-AHRQ).

Kravitz, RL. Park, RE. Kahan, JP (1997). 'Measuring the clinical consistency of panelists' appropriateness ratings: the case of coronary artery bypass surgery.' *Health Policy*, 42(2):135–143.

Lacourciere, J (ed.) (2007). *HEDIS 2007 technical specifications*. Washington, DC: National Committee for Quality Assurance (No. 2).

Leatherman, S. Ferris, TG. Berwick, D. Omaswa, FM, Crisp, N (forthcoming). *The role of quality improvement in strengthening health systems in developing countries*.

Lindenauer, PK. Remus, D. Roman, S. Rothburg, MB. Benjamin, EM. Ma, A. Bratzler, DW (2007). 'Public reporting and pay for performance in hospital quality improvement.' *New England Journal of Medicine*, 356(5): 486–496.

Mainz, J. Krog, BR. Bjornshave, B. Bartels, P (2004). 'Nationwide continuous quality improvement using clinical indicators: the Danish National Indicator Project.' *International Journal for Quality in Health Care*, 16(Suppl.1): 45–50.

Mattke, S. Kelley, E. Scherer, P. Hurst, J. Lapetra, MLG. HCQI Expert Group Members (2006). *Health care quality indicators project: initial indicators report*. Paris: OECD Publications Service (OECD Health Working Papers No. 22).

McGlynn, EA. Damberg, C. Kerr, E. Schuster, M (eds.) (2000). *Quality of care for children and adolescents: a review of selected clinical conditions and quality indicators*. Santa Monica, CA: RAND Corporation (Publication No. MR-1283-HCFA).

McGlynn, EA. Kerr, EA. Damberg, C. Asch, S (eds.) (2000a). *Quality of care for women: a review of selected clinical conditions and quality indicators*. Santa Monica, CA: RAND Corporation (Publication No. MR-1284-HCFA).

OECD (2007). *Health at a glance 2007: OECD indicators*. Paris: Organisation for Economic Co-operation and Development.

Ovretveit, J (2004). 'Formulating a health quality improvement strategy for a developing country.' *International Journal of Health Care Quality Assurance*, 17(7): 368–376.

Roland, M (2004). 'Linking physicians' pay to the quality of care – a major experiment in the United Kingdom.' *New England Journal of Medicine*, 351(14): 1448–1454.

Selby, JV. Fireman, BH. Lundstrom, RJ. Swain, BE. Truman, AF. Wong, CC. Froelicher, ES. Barron, HV. Hlatky MA (1996). 'Variation among hospitals in coronary-angiography practices and outcomes after myocardial infarction in a large health maintenance organization.' *New England Journal of Medicine*, 335(25): 1888–1896.

Shekelle, PG. Chassin, MR. Park, RE (1998). 'Assessing the predictive validity of the RAND/UCLA appropriateness method criteria for performing carotid endarterectomy.' *International Journal of Technology Assessment in Health Care*, 14(4): 707–727.

Shekelle, PG. Kahan, JP. Bernstein, SJ. Leape, LL. Kamberg, CJ. Park, RE (1998a). 'The reproducibility of a method to identify the overuse and underuse of medical procedures.' *New England Journal of Medicine*, 338(26): 1888–1895.

Shiffman, RN. Shekelle, P. Overhage, JM. Slutsky, J. Grimshaw, J. Deshpande, AM (2003). 'Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization.' *Annals of Internal Medicine*, 139(6): 493–498.

Steel, N. Melzer, D. Shekelle, PG. Wenger, NS. Forsyth, D. McWilliams, BC (2004). 'Developing quality indicators for older adults: transfer from the USA to the UK is feasible.' *Quality and Safety in Health Care*, 13(4): 260–264.

Wenger, N. Roth, CP. Shekelle, P. & ACOVE Investigators (2007). 'Introduction to the assessing care of vulnerable elders-3 quality indicator measurement set.' *Journal of the American Geriatrics Society*, 55(Suppl. 2): 247–52.

2.4 *Measuring financial protection in health*

ADAM WAGSTAFF

Introduction

Health systems are not just about improving health. Good ones also ensure that people are protected from the financial consequences of illness and death, or at least from the financial consequences associated with the use of medical care. Anecdotal evidence suggests that health systems often perform badly in this respect, with devastating consequences especially for poor and near-poor households. The World Bank participatory poverty study in fifty countries – *Voices of the Poor* (Narayan et al. 2000a) – found that poor health and illness are universally dreaded as a source of destitution, not only because of the costs of health care but also because of the income lost. The study documents the case of a twenty-six year-old Vietnamese man who was the richest man in his community but became one of the poorest as a result of the health-care costs incurred for his daughter's severe illness (Narayan 2000). Another case concerned a thirty year-old Indian mother of four who was forced to sell the family's home and land and must walk 10 km a day transporting wood on her head in order to finance the cost of her diabetic husband's medical care (Narayan 2000).

How can a health system's success in protecting people against the financial consequences of ill health be measured? What do successful systems have in common? How far do health system reforms improve people's financial protection against health expenses? This chapter provides an overview of the methods and issues arising in each case and presents empirical work on financial protection in health, including the impacts of government policy. The chapter also reviews a recent critique of the methods used to measure financial protection.

Some preliminaries

The measures of financial protection developed to date are based on out-of-pocket spending on medical care and relate these payments to

a threshold (Wagstaff & van Doorslaer 2003). The idea is that out-of-pocket spending is largely involuntary and does not contribute to household well-being in the way that spending on (say) a new car might. A household unfortunate enough to have to pay for medical care is deprived of resources that could be used to purchase other goods and services, including necessities such as food and shelter. One approach is to classify spending as catastrophic if it exceeds a certain fraction of household income. Another is to classify it as impoverishing if it is sufficiently large to make the difference to a household being above or below the poverty line, i.e. in the absence of the medical outlays the household's resources would have been sufficient to keep living standards above the poverty line; with the outlays living standards are pushed below the poverty line.

Three general issues arise with these approaches. First, the focus is the cost of medical care; income losses associated with illness, injury and death are not captured, even though they may have greater impacts on household welfare. The justification is that these measures aim to assess financial protection related to health-care expenses and that the social protection system should be responsible for protecting households against income losses. Second, the assumptions that out-of-pocket spending on health is involuntary and automatically deprives households of resources should be considered. They are discussed further below. Third, some argue that the focus on what households spend misses an important point – high out-of-pocket costs may deter some people from using health services. A country in which people pay little out of pocket (and which therefore looks good from a financial protection perspective) may be one in which people do not use health services. Some argue that this should be captured by a financial protection measure.

On the face of it, it seems reasonable that financial protection measures should capture forgone utilization caused by high out-of-pocket costs. However, this confuses policy objectives with policy instruments. Policy-makers seek to influence multiple (focal) variables including health outcomes and people's expenditure on health (and by implication their available resources for other goods and services). They have a number of instruments at their disposal, including the share of the cost of health care that people pay out of pocket. A change in a given instrument will likely affect several focal variables. For example, exempting poor people from user fees at public facilities will likely

affect their use of services (non-use and under-utilization should fall) and the amount that they pay out of pocket.

The natural approach to a health system assessment is to examine how the system functions in terms of the focal variables and works backwards to see how far this is attributable to specific set policies that have been adopted. For example, a country might show good financial protection but poor health outcomes and health inequalities if out-of-pocket payment policies discourage most people from using health services but those that are used are high quality and appropriate. Another country might have poor financial protection and poor health outcomes and inequalities because people use services (despite high cost at the point of use) that are poor quality or inappropriate for their needs. This example highlights that performance on financial protection depends not just on policies for narrowly defined health financing but also (amongst other things) on the way that providers are paid and regulated.

Catastrophic expenditures

The basics

Many studies simply examine the distribution of catastrophic health expenditures. These are defined as health spending that exceeds a threshold usually defined in relation to the household's pre-payment income. This is illustrated in Fig. 2.4.1 which plots out-of-pocket spending on medical care (M) against non-medical spending (NM) on other items such as food, housing, transport, etc. In Fig. 2.4.1 a household has income equal to x (intercept on x and y axes) and outgoings on medical care (M_0) and other items (NM_0). The 45° budget line indicates that each dollar spent on medical care means one dollar less to spend on other things. It is this fact that underpins the concern over financial protection – that medical care outlays are different from spending on other goods and services. They are viewed as involuntary responses to unwanted health shocks and are considered to have entirely negative effects on households by diverting resources that could have been spent on goods and services that contribute to welfare. Waters et al. (2004) define out-of-pocket medical spending as catastrophic if it exceeds a certain amount.

Wagstaff and van Doorslaer (2003), by contrast, consider spending is catastrophic if it exceeds some specified fraction of pre-payment income (x) defined as the sum of observed medical outlays (M_0) and observed

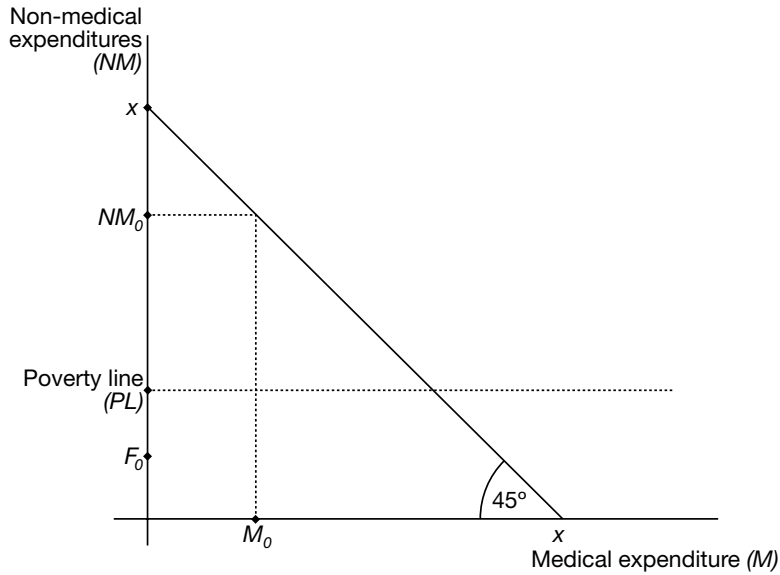

Fig. 2.4.1 Defining catastrophic health spending

Source: Author's own compilation

non-medical spending (NM_0). The threshold could also be defined in terms of pre-payment income less a deduction for food and (possibly) other necessities (Wagstaff & van Doorslaer 2003; Xu et al. 2003). The idea is that these deductions for basic necessities offer a better idea of an individual's ability to pay. These deductions could be an individual's (or household's) actual food expenditure (F_0) or what is considered to be the minimum acceptable level of expenditure on food (and perhaps other necessities) as reflected in a poverty line (PL). The latter approach is problematic when a household's pre-payment income falls short of the poverty line. In such cases, households have a negative estimated ability to pay that automatically falls below the catastrophe threshold whatever the medical care outlay (Wagstaff & van Doorslaer 2003).[1]

[1] Xu et al. (2003) use this approach. Their poverty line is just for food expenditures, which is subtracted from non-medical consumption (NM_0) rather than pre-payment income (x). Ability to pay is defined as NM_0-PL except for households for whom this is negative. In such cases, ability to pay is defined as NM_0 less *actual* food expenditure. This leads to the rather unsatisfactory outcome that a household just below their poverty line could be judged to have the same ability to pay as one just above it.

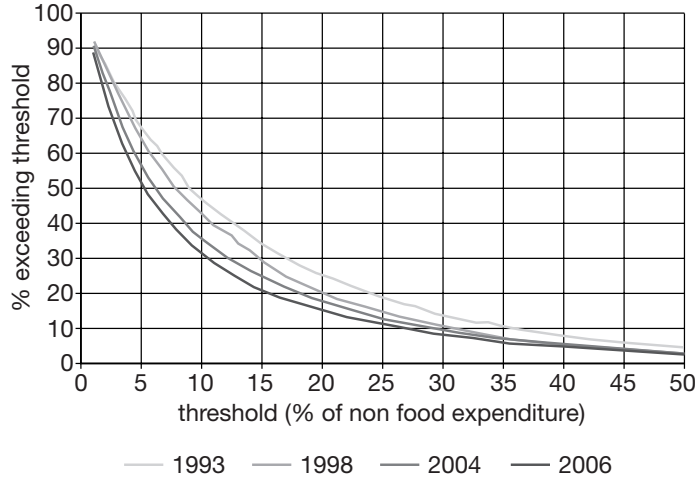

Fig. 2.4.2 Catastrophic spending curves, Viet Nam

Source: Author

Of course, the precise fraction of pre-payment income (with or without some deduction for basic necessities) is arbitrary; therefore it is sensible to examine the sensitivity of results to the threshold chosen. Fig. 2.4.2 shows catastrophic spending curves for a variety of years for Viet Nam – plotting the fraction of households experiencing catastrophic out-of-pocket spending (y-axis) for a given threshold (x-axis). In this instance, the incidence of catastrophic spending has fallen continuously over the period, whatever the threshold, and therefore the choice of threshold is irrelevant.

It may be desirable to move beyond counting the number of households who overshoot the threshold to capturing the amount by which they overshoot it. This is common in the poverty literature which assesses not only the number of people in poverty but also the poverty gap – the extent to which they fall below the poverty line. The catastrophic payment gap is simply the aggregate or average amount by which out-of-pocket spending exceeds the threshold (Wagstaff & van Doorslaer 2003). Fig. 2.4.3 plots out-of-pocket payments as a share of income (y-axis) against the cumulative share of the population (x-axis), ranked in decreasing order of out-of-pocket payments as a share of income. The catastrophic payment headcount (those whose payments exceed the threshold) is obtained by reading off the curve

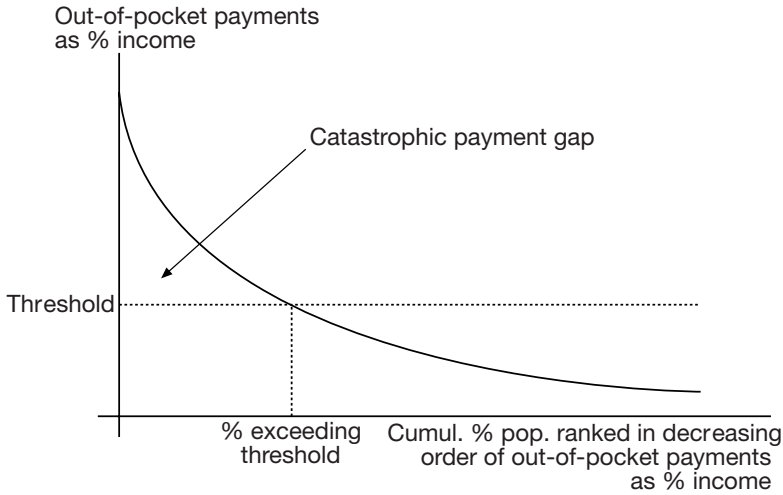

Fig. 2.4.3 Catastrophic spending gap

Source: Wagstaff & van Doorslaer 2003.

at the threshold. The (aggregate) catastrophic payment gap shows the overall amount by which payments exceed the threshold in the sample.

A final modification is to make some allowance for whether well-off or worse-off households exceed the threshold. It is likely that policymakers would be more concerned about the latter. The incidence of catastrophic payments and the catastrophic payment gap could be tabulated by pre-payment income quintile or by computing a concentration index for each (Wagstaff & van Doorslaer 2003). For example, the concentration index for the catastrophic health expenditure headcount would be negative if catastrophic expenditures were, on average, more common among the worse off. Of course, it may be that the fraction of the population experiencing catastrophic spending has increased over time but become less concentrated among the poor. Multiplying the catastrophic payment headcount by the complement of the concentration index provides a natural summary measure that takes both into account (Wagstaff & van Doorslaer 2003). This is equivalent to constructing a rank-weighted average of the binary variable indicating whether or not the person in question had expenses exceeding the catastrophic payment threshold, where the weight is decreasing in the person's rank in the income distribution.

Empirical studies

Xu et al. (2003) found large differences when they reported the incidence of catastrophic health spending (using a 40% threshold) in fifty-nine countries (Fig. 2.4.4). Xu et al. (2007) recently produced estimates for eighty-nine countries covering 89% of the world's population, again using the 40% threshold. Their estimates range from 0% in the Czech Republic, Slovakia and the United Kingdom to more than 10% in Brazil and Viet Nam. Several OECD countries (Portugal, Spain, Switzerland, United States) record rates in excess of 0.5%.

Van Doorslaer et al. (2007) looked at catastrophic spending in ten Asian territories. They found relatively low rates in Malaysia, Sri Lanka and Thailand and relatively high rates in China, Viet Nam and Bangladesh. This study also looked at the pre-payment income distribution of those experiencing catastrophic payments. For the most part, catastrophic spending was concentrated among the better off although this was dependent to some degree on the threshold chosen. Taiwan is the exception – catastrophic spending was concentrated among the poor whatever the threshold. A different picture emerges in Waters et al's (2004) study in the United States. They found a higher incidence of catastrophic spending among poor families and those with multiple chronic conditions. In Belgium, too, the incidence was found to be higher among poorer families (De Graeve & Van Ourti 2003).

A number of studies explore how policies and institutions impact on the incidence of catastrophic health spending. Xu et al. (2003 & 2007) found that rates of catastrophic spending are higher in poorer countries and in those with limited prepayment systems. Xu et al's (2007) most recent study (controlling for whether prepayment as a share of health spending exceeds 50%) found that the incidence of catastrophic spending does not vary between tax-financed or social health insurance systems. Looking at cross-country differences, van Doorslaer et al. (2007) speculate that the low incidence of catastrophic spending in Sri Lanka, Malaysia and Thailand reflects the low reliance on out-of-pocket spending to finance health care and the limited use of user fees in the public sector. By contrast, the high rate of incidence in the Republic of Korea is argued to reflect the high copayments in that country's social insurance system and the partial coverage of inpatient care. De Graeve and Van Ourti (2003) found that the incidence of catastrophic spending in

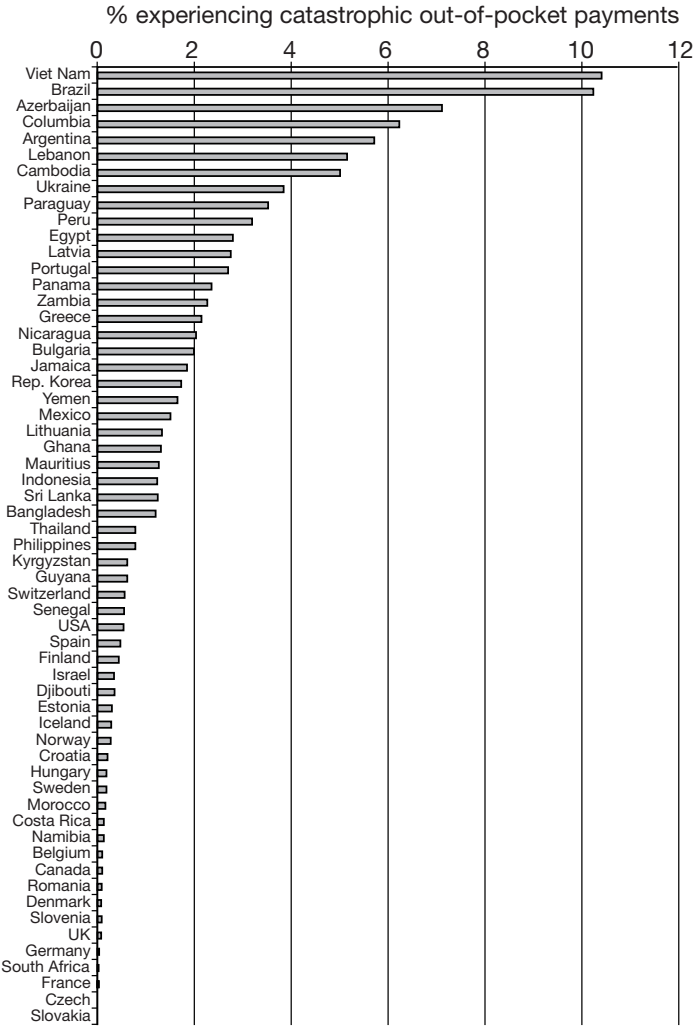

Fig. 2.4.4 Incidence of catastrophic out-of-pocket payments in fifty-nine countries

Source: Xu et al. 2003

Belgium would have been higher without a policy that imposes a ceiling on official out-of-pocket payments linked to a family's income. This ceiling has greatest effect in the middle of the income distribution.

Several country-level studies conclude that insurance reduces the risk of catastrophic health spending. Gakidou et al. (2006) and Knaul et al. (2006) found that the introduction of the Popular Health Insurance scheme in Mexico from 2001 led to a reduction in the incidence of catastrophic health expenditures. Limwattananon et al. (2007) found that rates of catastrophic spending in Thailand were lower after the universal health-care scheme was introduced in 2001. Habicht et al. (2006) found that the risk of catastrophic spending in Estonia increased during the late 1990s and early 2000s. They attribute this partly to rising copayments (hence a decrease in the depth of coverage) linked to a decline (in real terms) in government health spending and partly to the ageing of the population – elderly people have shallower coverage, especially for medicines.

Other studies point to the limitations of using insurance to reduce and eliminate catastrophic spending. Wagstaff and Pradhan (2005) found that the introduction of a social health insurance scheme in Viet Nam in 1993 reduced the incidence of catastrophic expenses. Wagstaff (2007) found that the scheme's subsequent extension to the poor (financed through general revenues) produced similar results. However, the percentage reductions were estimated to be small and high rates of catastrophic spending were observed even among those with insurance. These results may be explained partly by the fact that insurance appears to have increased the utilization of services in Viet Nam. Xu et al. (2006) found lower rates of catastrophic out-of-pocket spending among the Ugandan population following the removal of user fees in 2001 although the rate increased among the poor. They speculate that this was due to the frequent unavailability of drugs at government facilities following the removal of user fees – patients were forced to buy drugs from private pharmacies and informal payments to health workers increased to offset lost revenues from fees. Devadasan et al. (2007) examined how two community health insurance schemes in India affected the risk of catastrophic out-of-pocket payments and concluded that they halved the risk. This limited impact on benefit packages is attributed to low maximum limits; the exclusion of some conditions from the package; and the use of the private sector for some inpatient admissions.

Ekman (2007) found that insurance increases the risk of catastrophic spending in Zambia and suggests that the amount of care per illness episodes may have increased. He contends that quality assurance and the oversight of service providers is important in determining how far insurance reduces the risk of catastrophic spending. Three recent studies from China reinforce these points. Wagstaff and Lindelow (2008) found that China's urban insurance scheme increases the risk of catastrophic out-of-pocket spending. These results are attributed in part to weak regulation of providers; a fee-for-service payments system; and a fee schedule that allows providers to profit from drugs and the high-tech care results for insured patients receiving more complex care and from higher-level (hence more costly) providers. Wagstaff et al. (2007) found that China's new rural insurance scheme does not appear to have reduced the incidence of catastrophic health spending. They attribute this to exclusions, high deductibles, low reimbursement ceilings and similar supply responses to those seen in the urban setting. By contrast, Wagstaff and Yu (2007) found that supply-side interventions in rural China (including the introduction of treatment protocols and essential drug lists) reduced the incidence of catastrophic health spending.

Impoverishing expenditures

The basics

The catastrophic payment approach is limited by its failure to show the extent of the hardship caused by catastrophic payments. One household might spend more than 25% of its pre-payment income on health and yet be nowhere near the poverty line. Another might spend only 1% of its pre-payment income before crossing the poverty line. Impoverishment offers an alternative perspective – the core idea being that health-care expenses should push no one into (or further into) poverty.

A household may be classified as impoverished by out-of-pocket payments on medical care if pre-payment income (x in Fig. 2.4.1) lies above the poverty line (PL) and non-medical spending (NM_0) lies below (Wagstaff & van Doorslaer 2003). Comparison of the pre-payment poverty headcount (fraction of households where $x>PL$) and the post-payment poverty headcount (fraction of households where

NM_0<PL) can indicate how far out-of-pocket payments cause impoverishment by identifying the fraction of the population that crosses the poverty line as a result of health expenditures. This approach does not capture how far people are pushed below the poverty line as a result of health spending or the possibility that health spending may push already poor households (in terms of their pre-payment discretionary income) into greater poverty. This can be established by comparing the pre-payment poverty gap (aggregate shortfall from poverty line using x as the living standards measure) with the post-payment poverty gap (aggregate shortfall from poverty line using NM_0 as the living standards measure).

Empirical studies

Wagstaff and van Doorslaer (2003) looked at health-care payments and poverty in Viet Nam in 1993 and 1998. Fig. 2.4.5 shows their pre-payment income Pen's parade for Viet Nam in 1998. This paint drip chart also shows households' out-of-pocket payments and a food-based poverty line. The difference between the pre-payment and post-payment poverty headcount is around 3.5% and the difference between the pre-payment and post-payment (normalized) poverty gaps is around 1%. In 1993, the difference between the pre-payment and post-payment poverty headcounts was 4.4%. This greater fall in the headcount for post-payment income reflects the fall in the share of income absorbed by health spending over this period in Viet Nam (Wagstaff 2002).

Results for rural China over the same period show a reduction in the difference between pre-payment and post-payment headcounts (Liu et al. 2003). However, Gustafsson and Li (2004) found the opposite in their analysis of changes between 1988 and 1995. The poverty headcount fell by 2.2% at the dollar-a-day poverty line when health expenditures were not deducted from disposable income; and by only 0.7% percentage points when they were. This reflects the fact that the share of income spent on health care increased in rural China during the period 1988–1995.

Two studies have looked at trends before and after the introduction of a reform. Limwattananon et al. (2007) found that impoverishment rates in Thailand were lower (but not zero) following the introduction of the universal health-care scheme in 2001. The failure to eliminate

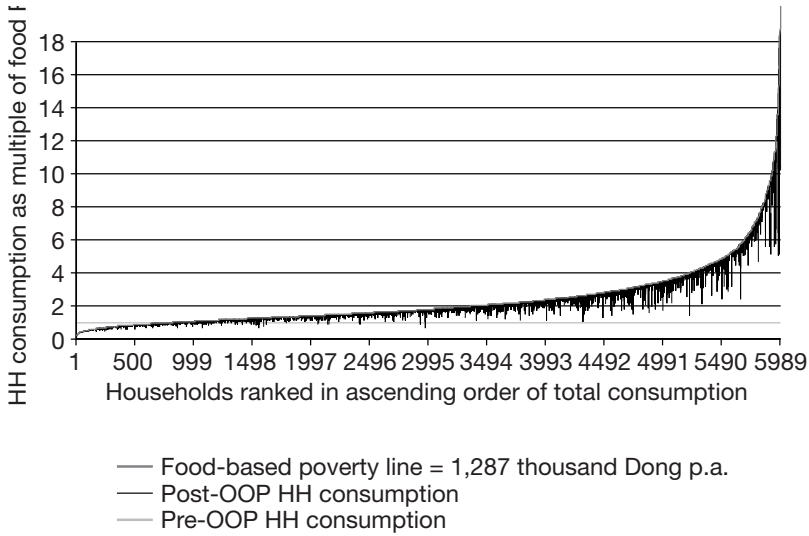

Fig. 2.4.5 Out-of-pocket payments and poverty in Viet Nam, 1998

Source: Wagstaff & van Doorslaer 2003.

impoverishment caused by out-of-pocket expenses is attributed to people who bypass their designated provider and thus make themselves unnecessarily liable for out-of-pocket payments and non-coverage of certain interventions, e.g. renal dialysis and chemotherapy. Knaul et al. (2006) report that the difference between the pre-payment and post-payment poverty gap narrowed following the introduction of the Popular Health Insurance scheme in Mexico.

Van Doorslaer et al. (2006) used data from eleven Asian countries to compare pre- and post-payment poverty headcounts and poverty gaps using the World Bank's dollar-a-day poverty line (as well as its US$ 2-a-day poverty line). On average, they found that the dollar-a-day poverty headcount is almost three percentage points higher when out-of-pocket spending is deducted from household consumption. The difference is almost four percentage points in Bangladesh and India but just 0.1 and 0.3 percentage points in Malaysia and Sri Lanka, respectively.

Alam et al. (2005) compared pre-payment and post-payment poverty headcounts in ten countries in eastern Europe and the former Soviet Union using a US$ 2.15-a-day poverty line at 2000 prices and purchasing power parities. On average, out-of-pocket payments raise

the poverty headcount by 2% percentage points – Armenia (3.4), Georgia (3.6) and Tajikistan (3.3) recorded the highest increases. Interestingly, the average share of income spent on out-of-pocket health care payments is quite different in Armenia (around 12%) and Georgia (around 7%). However, the shares among the poorest and second poorest quintiles are quite similar at around 14% and 8%, respectively. The high incidence of impoverishment due to health-care spending in these countries likely reflects the collapse of publicly-financed health systems and increasing reliance on out-of-pocket payments, including informal ones. The rate in Armenia would probably have been even higher if the government's 2001 reform had not provided the services in the health insurance scheme's benefit package free of charge to households receiving social assistance.

Is health spending involuntary?

The catastrophe and impoverishment approaches outlined above make two key assumptions. The first is that health-care payments should be seen as involuntary and non-discretionary – the result of an unforeseen and unwanted shock and rarely the result of a deliberate choice by the individual concerned. In this view, health-care payments stand apart from other items of household consumption that contribute to household welfare or utility.

This view can be challenged as in some cases individuals may well have some discretion (at least at the margin) over health expenditures. However, generally it seems more reasonable to treat health spending as non-discretionary and to consider that it does not contribute to household welfare. This would exclude it from household spending in consumption aggregates used in studies of household living standards. Deaton and Zaidi (1999) reached a similar conclusion based partly on the low income elasticities of health spending they found in six of the seven developing countries they studied. Burtless and Siegel (2001) also argue for this approach in their discussion of proposals to take explicit account of health-care spending when computing poverty rates in the United States.

It seems reasonable to treat health expenditures as involuntary but the implied practice of excluding out-of-pocket spending from consumption aggregates for measuring poverty is often not followed. For example, the World Bank's official dollar-a-day poverty figures

are based on measures of household consumption that include out-of-pocket spending on medical care. This produces poverty rates that are lower than they would be if out-of-pocket spending on medical care was treated as involuntary and excluded from the consumption aggregate (van Doorslaer et al. 2006).

Asset sales, dissaving and borrowing

The second assumption that underpins the catastrophe and impoverishment approaches is that a household's non-medical expenditure in the period under consideration would have increased by an amount equal to its out-of-pocket expenditures on medical care had it not incurred the out-of-pocket spending. In other words, it is assumed that the household was forced to finance the health spending entirely from its current non-medical consumption.

This assumption fails if the household is able to finance some (or all) of the expenditure by running down its stock of financial and physical assets (dissaving) or by borrowing. In both cases, current income (gross of proceeds of asset disposals and loans taken) is higher when medical costs are incurred than when they are not. Fig. 2.4.6 illustrates a household that spends M_0 on medical care and NM_0 on other things. If the household member needing medical care had not fallen ill, the household's income would have been x' not x. The difference between the two reflects the proceeds of asset sales or funds from a gift or loan. The drop in non-medical consumption caused by the use of medical care (ultimately the quantity of interest) is equal to the difference between x' and NM_0. This is less than out-of-pocket spending (M_0) in cases such as that illustrated in Fig. 2.4.6 when people are able to borrow or sell assets to reduce the impact of health spending on non-medical consumption. Indeed, it may well be that the household is completely able to smooth its non-medical consumption in the face of health shocks that necessitate health expenditure. In the case illustrated, x' and NM_0 coincide and the medical expenses cause no reduction in non-medical consumption. The household is only partially able to smooth non-medical consumption in the face of health shocks and non-medical consumption is cut back in the period when the health shock occurs. However, this reduction is less than the amount of the medical expenditure. The reduction in non-medical consumption equals the amount of health expenditures only in extreme cases when

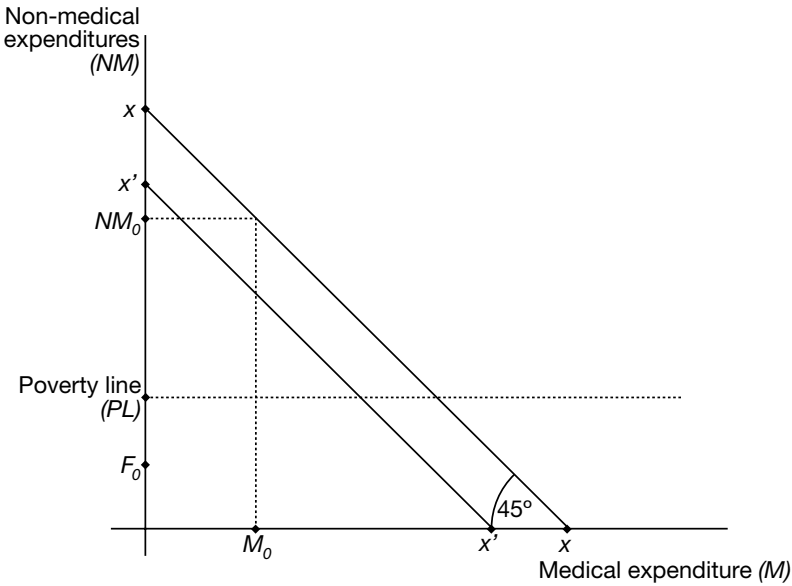

Fig. 2.4.6 Case where health spending is not financed out of current income
Source: Author

the household is unable to use savings or borrow, as illustrated in Fig. 2.4.1.

Empirical evidence suggests that people do prevent drops in non-medical consumption by selling assets or borrowing. The World Health Survey (WHS) asked how people finance their health expenditures (http://www.who.int/healthinfo/survey/instruments/en/index.html). Respondents were able to choose from the following sources: savings; selling items; borrowing from relatives; borrowing from others; health insurance; current income; and other. Fig. 2.4.7 shows the cumulative percentages for a selection of countries; the y-axis would have been 700% if people had used all seven sources. It seems likely that people in countries with pre-payment schemes financed from general revenues and no out-of-pocket payments would select none of the seven options. These people are unlikely to consider that the pre-payment scheme is insurance. This explains why South Africa and Sri Lanka average less than 100%. The clear message from Fig. 2.4.7 and from other surveys is that people borrow, sell assets and dissave to protect their living standards in the face of health shocks that necessitate out-

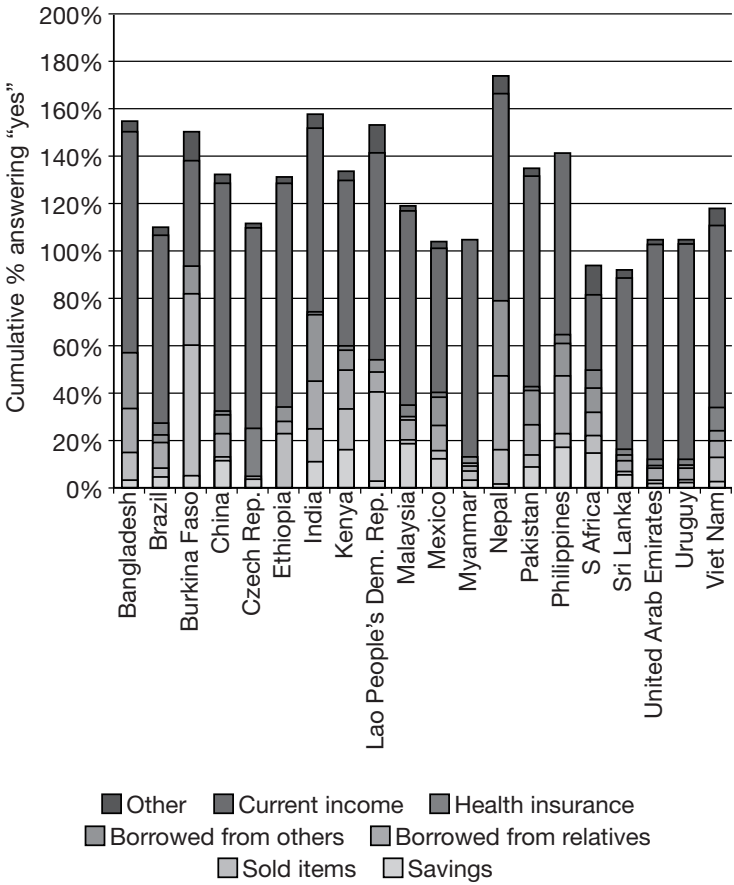

Fig. 2.4.7 How households finance their health spending, selected countries

Source: World Health Surveys (http://www.who.int/healthinfo/survey/whsresults/en/index.html)

of-pocket spending on care. The mix of strategies varies from country to country. Countries where asset disposals feature prominently are likely be those in which households find it difficult to get credit.

Whatever the sources used to protect living standards in the face of health shocks, it is important to allow for such strategies when estimating people's financial protection against health expenditures. Failure to do so will result in an overestimate of the extent to which health expenditures are catastrophic and impoverishing and an underestimate of the related degree of financial protection (provided by one method

or another). As far as catastrophic spending is concerned, the numerator in Fig. 2.4.6 (originally M_0) should be replaced by the drop in non-medical consumption caused by the medical expenditure (x'-NM_0) and the denominator (x) should be replaced by the amount of non-medical consumption that would have been enjoyed in the absence of the health shock (x'). For impoverishment, the pre-payment headcount should be assessed on the basis of x', rather than x, and the post-payment poverty headcount computed using observed non-medical consumption ($NM_{0)}$. Further doubt is raised about including out-of-pocket payments in the consumption aggregate for measuring poverty when dissaving, asset sales and borrowing are factored in (van Doorslaer et al. 2006). Medical outlays financed largely by dissaving and borrowing may push a household above the poverty line when non-medical and medical expenditure is combined. A health financing reform that cuts out-of-pocket payments and reduces the need for households to dissave and borrow would actually increase measured poverty.

Modification of estimates of catastrophe and impoverishment to take account of dissaving, asset sales and borrowing requires an estimate of the counterfactual income (x') – a household's income in the absence of the health expenditures. The WHS is one of the few household surveys to ask how households financed their health expenditures. Questions about what was raised by selling assets or borrowing are asked sometimes in specialized vulnerability surveys but rarely in health surveys. The 1995 Indian National Sample Survey is an exception. In their analysis of the data, Flores et al. (2008) found heavy use of coping strategies including drawing down of savings, asset sales, borrowing and transfers. They found that such strategies finance three quarters of the cost of inpatient care in rural areas and two thirds in urban areas. They also find that these sources fully finance hospital costs in 52% of rural and 44% of urban households. Ignoring the use of coping strategies to protect current income suggests that 2.2% of rural Indian households incur catastrophic payments for inpatient care using a 5% threshold. This estimate is reduced to just 0.2% following the adjustments outlined above.

Flores et al. (2008) found similar dramatic differences for impoverishment in urban households. In rural areas, the poverty headcount for actual non-medical consumption (NM_0 in Fig. 2.4.6) is 39.45% and the headcount corresponding to the naive estimate of what non-medical consumption would have been in the absence of medical outlays

(x) is 8.94%. The naive approach would indicate that out-of-pocket payments have raised poverty dramatically. However, the headcount for what non-medical consumption would have been in the absence of medical outlays and factoring in people's coping strategies (x') is just 39.39%, barely different from the actual poverty rate.

These results indicate that households are generally able to smooth non-medical consumption in the face of large outlays on medical care. This appears at odds with the econometric literature that looks at the effects of health shocks on household non-medical consumption. Typically, that literature finds that households are unable to smooth consumption in the face of health shocks, at least large ones (Gertler & Gruber 2002; Wagstaff 2007). However, outlays on medical care are just one channel through which health shocks affect non-medical consumption. Losses in earned income (possibly offset at least in part by increases in unearned income) are another, possibly more important, channel and evidence suggests that households are unable to smooth consumption in the face of income shocks (Jalan & Ravallion 1999). Therefore, the two literatures are, in fact, not at odds with one another.

Intertemporal considerations

Flores et al. (2008) acknowledge that the argument in the previous section misses the fact that households incur costs to finance out-of-pocket payments. These costs should not be disregarded when measuring catastrophic and impoverishing payments. Households with insurance cover for out-of-pocket payments likely have reduced uncertainty about future expenditures and are able to hold their wealth in less liquid forms that offer higher returns. In addition, loans have to be repaid (possibly at very high interest rates) in subsequent periods and returns on assets and savings are lost when these have been sold or used.

Flores et al. (2008) provide an example of an Indian high-spending household in which per capita consumption is INR 6866 and inpatient out-of-pocket payments are INR 2760. The household finances these payments by borrowing INR 1020; drawing INR 823 from savings; and raising INR 298 from asset sales, INR 439 from other sources and INR 180 from current income. Flores et al. (2008) focus on the INR 180 financed from current income and ignore the other expenses.

They compute the coping-adjusted expense ratio by dividing 180 (sum financed by current income) by 4286 (6866 consumption less 2580 out-of-pocket payments financed through coping strategies). This is just 4%, one tenth of the conventional ratio of out-of-pocket spending divided by consumption (2760/6866=40%). Even for the current period, 4% is likely to underestimate the hardship caused by medical care costs – forgone returns accrue from the moment that assets and savings are cashed in and loan repayments are likely to start well within twelve months of the expenses being incurred. In any case, costs incurred beyond the current period should not simply be ignored.

What might the time path of expenses look like for this Indian household? Banerjee and Duflo (2007) report monthly interest rates of 3%–4% among India's poor. If the INR 1020 loan was repaid over three years at a 3.5% monthly interest rate then the household's annual repayments would be INR 607. Suppose that in the absence of medical-care expenses the household would have held the savings and assets for three years. If the INR 823 of savings and the INR 298 of assets earned 10% per annum then, on average, they would have produced combined annual returns of INR 129. Loan interest and lost returns give a total cost of INR 736 for each of the three years following the inpatient expenditure. This can be compared with the household's per capita consumption in the absence of the interest payments and forgone returns – INR 4842 (6866-2760+736). The ratio of 736 to 4842 is 15%, considerably less than the 40% produced by the naive calculation but a good deal higher than the 4% from the calculation above. For some thresholds this might be considered catastrophic. Obviously these calculations hinge on assumptions about the duration of the loan; loan interest rates; the number of years the assets and savings would have been held in the absence of the shock; and the interest that the household would have earned on them.

The example above provides a somewhat truer picture but still misses something. It overlooks the fact that households are likely to incur at least some medical outlays every year – possibly even quite high costs for several years in a row. So, while it is true that a health shock in year t may not cause a major drop in consumption in that year (if any) because the household borrows to finance the cost of medical care, it is also possible that the household may already be paying off a loan for a previous health shock in year t-2. This is more

likely to be the case if health expenditures are highly correlated across years at the household level, i.e. if households that incur expenditures in one year are more likely to incur expenditures in subsequent years.

The rank correlation for health expenditures over the five years between the two waves of the Viet Nam 1993-98 Living Standards Measurement Study panel is 0.36. This is lower than the rank correlation for non-medical consumption (0.66) but still quite high.[2] Over the two years between the two waves of the China panel used by Wagstaff et al. (2007) the rank correlation for medical outlays at household level is 0.31, compared to 0.66 for household income. With correlations of this size, episodes of coping with expenses incurred following health shocks will likely overlap. In the example of the Indian household given above this might require the estimated interest payments and forgone returns for the INR 2760 medical bill to be added to similar charges incurred earlier. Thus, Flores et al's (2008) 15% figure is likely to be an underestimate of the hardship caused by medical bills, possibly a considerable underestimate.

Conclusions

There has been a good deal of progress in designing and implementing measures of financial protection in health but, perhaps inevitably, the work is incomplete. One major challenge concerns how to take account of how people finance their medical outlays and when they incur the costs. The recent literature (Flores et al. 2008) is right to reiterate that, contrary to what is assumed by the naive approach used to date, households may not experience much of a drop in living standards during the period in which the outlays are made. However, households do have to make sacrifices at some stage. Borrowing allows the sacrifice to be deferred and spread over multiple periods, although interest rates will add to the bill. Furthermore, households are unlikely to incur out-of-pocket payments on a one-off basis and more likely to incur at least some expenses every year. A household may have to borrow to finance a medical care bill precisely because it has not yet repaid the loan that financed earlier charges. The challenge is to move from the snapshot approach that assumes that outlays entail

[2] Author's calculation from the Viet Nam 1993 and 1998 Living Standards Measurement Study data.

consumption sacrifices in the period in which they are incurred to an intertemporal approach that takes account of the (possibly quite different) time paths of outlays and forgone consumption.

The naive approach assumes that consumption drops pari passu with medical outlays and is therefore likely to underestimate the hardship caused by out-of-pocket spending. However, it remains useful as it has the merit of capturing the amount of money that households must find (one way or another) and relating this to their standard of living. Furthermore, it can be implemented with a standard household expenditure or multipurpose survey. By contrast, the alternative approach focuses (purportedly) on costs incurred in the current period and ignores those incurred in other periods. For this reason, and because it overlooks the fact that some costs (e.g. forgone returns on assets and loan repayments) are likely to be incurred in the period in which the medical bills are incurred, it is likely to provide a lower, possibly highly conservative, bound.

Subject to the caveats associated with the naive methods of measuring financial protection, some general points emerge from the empirical literature. Financial protection in health appears to vary across countries, partly reflecting the role of per capita income. On average, higher rates of catastrophic payments are found in poorer countries and therefore those who can least afford large out-of-pocket payments for health care are at greatest risk. However, differences exist across countries at a given per capita income. These appear to reflect income inequality and also the extent to which health-care payments are prepaid through some form of insurance.

The roles of insurance, pre-payment and other forms of financial protection emerge from country studies. Expansion of insurance coverage tends to reduce the incidence of catastrophic spending and impoverishment, while a reduction in the depth of coverage has tended to be associated with higher rates. As expected, ceilings on out-of-pocket payments reduce the incidence of catastrophic spending. But there are caveats. Studies point to a variety of factors that together influence the degree to which insurance influences financial protection.

- Insurance tends to increase the quantity of care received and puts upward pressure on out-of-pocket payments.
- Some benefit packages are not especially generous, with high deductibles, high coinsurance rates, low reimbursement ceilings

and multiple exclusions (for example, drugs which often use a large share of household health spending).

- Providers may not be properly compensated by third-party payers. They may look to informal payments to make up lost income and may be unable to procure drugs on the terms offered by the third-party payer.

In China, recent research suggests that supply-side interventions (treatment protocols, drug lists, and so on) have had more success in improving financial protection than expansion of insurance coverage. This reinforces the point made earlier in this chapter – policy-makers have a variety of instruments available to increase financial protection in health. Insurance coverage is just one important instrument and it may not be the most effective for all applications.

References

Alam, A. Murthi, M. Yemtsov, R. Murrugarra, E. Dudwick, N. Hamilton, E. Tiongson, E (2005). *Growth, poverty, and inequality: eastern Europe and the former Soviet Union*. Washington, DC: World Bank.

Banerjee, AV. Duflo, E (2007). 'The economic lives of the poor.' *Journal of Economic Perspectives*, 21(1): 141–167.

Burtless, G. Siegel, S (2001). *Medical spending, health insurance, and measurement of American poverty*. Washington, DC: Brookings Institution (CSED Working Paper No. 22).

Deaton, A. Zaidi, S (1999). *Guidelines for constructing consumption aggregates for welfare analysis*. Princeton, NJ: Princeton University Woodrow Wilson School of Public and International Affairs (Research Program in Development Studies Working Paper No. 192).

De Graeve, D. van Ourti, T (2003). 'The distributional impact of health financing in Europe: a review.' *World Economy*, 26(10): 1459–1479.

Devadasan, N. Criel, B. Van Damme, W. Ranson, K. Van der Stuyft, P (2007). 'Indian community health insurance schemes provide partial protection against catastrophic health expenditure.' *BMC Health Services Research*, 7: 43.

Ekman, B (2007). 'Catastrophic health payments and health insurance: some counterintuitive evidence from one low-income country.' *Health Policy*, 83(2–3): 304–313.

Flores, G. Krishnakumar, J. O Donnell, O. van Doorslaer, E (2008). 'Coping with health-care costs: implications for the measurement of

catastrophic expenditures and poverty.' *Health Economics*, 17(12): 1393–1412.

Gakidou, E. Lozano, R. Gonzalez-Pier, E. Abbott-Klafter, J. Barofsky, JT. Bryson-Cahn, C. Feehan, DM. Lee, DK. Hernandez-Llamas, H. Murray, CJL (2006). 'Health system reform in Mexico 5 – assessing the effect of the 2001–06 Mexican health reform: an interim report card.' *Lancet*, 368(9550): 1920–1935.

Gertler, P. Gruber, J (2002). 'Insuring consumption against illness.' *American Economic Review*, 92(1): 51–76.

Gustafsson, B. Li, S (2004). 'Expenditures on education and health care and poverty in rural China'. *China Economic Review*, 15(3): 292–301.

Habicht, J. Xu, K. Couffinhal, A. Kutzin, J (2006). 'Detecting changes in financial protection: creating evidence for policy in Estonia.' *Health Policy and Planning*, 21(6): 421–431.

Jalan, J. Ravallion, M (1999). 'Are the poor less well insured? Evidence on vulnerability to income risk in rural China.' *Journal of Development Economics,* 58(1): 61–81.

Knaul, FM. Arreola-Ornelas, H. Mendez-Carniado, O. Bryson-Cahn, C. Barofsky, J. Maguire, R. Miranda, M. Sesma, S (2006). 'Health system reform in Mexico 4. Evidence is good for your health system: policy reform to remedy catastrophic and impoverishing health spending in Mexico.' *Lancet*, 368(9549): 1828–1841.

Limwattananon, S. Tangcharoensathien, V. Prakongsai, P (2007). 'Catastrophic and poverty impacts of health payments: results from national household surveys in Thailand.' *Bulletin of the World Health Organization*, 85(8): 600–606.

Liu, Y. Rao, K. Hsiao, WC (2003). 'Medical expenditure and rural impoverishment in China.' *Journal of Health Population and Nutrition*, 21(3): 216–222.

Narayan, D. Chambers, R. Shah, MK. Petesch, P (2000). *Voices of the poor: crying out for change*. New York, NY: Oxford University Press.

Narayan, D. Patel, R. Schafft, K. Rademacher, A. Koch-Schulte, S (2000a). *Voices of the poor: can anyone hear us?* New York, NY: Oxford University Press.

Van Doorslaer, E. O Donnell, O. Rannan-Eliya, RP. Somanathan, A. Adhikari, SR. Garg, CC. Harbianto, D. Herrin, AN. Huq, MN. Ibragimova, S. Karan, A. Ng, CW. Pande, BR. Racelis, R. Tao, S. Tin, K. Tisayaticom, K. Trisnantoro, L. Vasavid, C. Zhao, Y (2006). 'Effect of payments for health care on poverty estimates in 11 countries in Asia: an analysis of household survey data.' *Lancet*, 368(9544): 1357–1364.

Van Doorslaer, E. O Donnell, O. Rannan-Eliya, RP. Somanathan, A. Adhikari, SR. Garg, CC. Harbianto, D. Herrin, AN. Huq, MN. Ibragimova, S.

Karan, A. Lee, TJ. Leung, GM. Lu, JF. Ng, CW. Pande, BR. Racelis, R. Tao, S. Tin, K. Tisayaticom, K. Trisnantoro, L. Visasvid, C. Zhao, Y (2007). 'Catastrophic payments for health care in Asia'. *Health Economics*, 16(11): 1159–1184.

Wagstaff, A (2002). 'Reflections on and alternatives to WHO's fairness of financial contribution index.' *Health Economics*, 11(2): 103–115.

Wagstaff, A (2007). *Health insurance for the poor: initial impacts of Vietnam's health care fund for the poor*. Washington, DC: World Bank (Impact Evaluation Series no. 11. Policy Research Working Paper No. WPS 4134).

Wagstaff, A (2007a). 'The economic consequences of health shocks: evidence from Vietnam.' *Journal of Health Economics*, 26(1): 82–100.

Wagstaff, A. Lindelow, M (2008). 'Can insurance increase financial risk? The curious case of health insurance in China.' *Journal of Health Economics*, 27(4): 990–1005.

Wagstaff, A. Pradhan, M (2005). *Health insurance impacts on health and nonmedical consumption in a developing country*. Washington, DC: World Bank (Policy Research Working Paper No. 3563).

Wagstaff, A. van Doorslaer, E (2003). 'Catastrophe and impoverishment in paying for health care: with applications to Vietnam 1993–1998.' *Health Economics*, 12(11): 921–934.

Wagstaff, A. Yu, S (2007). 'Do health sector reforms have their intended impacts? The World Bank's Health VIII project in Gansu province, China.' *Journal of Health Economics*, 26(3): 505–535.

Wagstaff, A. Lindelow, M. Gao, J. Xu, L. Qian, J (2007). *Extending health insurance to the rural population: an impact evaluation of China's new cooperative medical scheme*. Washington, DC: World Bank (Impact Evaluation Series no. 12. Policy Research Working Paper No. WPS 4150).

Waters, HR. Anderson, GF. Mays, J (2004). 'Measuring financial protection in health in the United States.' *Health Policy*, 69(3): 339–349.

Xu, K. Evans, DB. Kawabata, K. Zeramdini, R. Klavus, J. Murray, CJ (2003). 'Household catastrophic health expenditure: a multicountry analysis.' *Lancet*, 362(9378): 111–117.

Xu, K. Evans, DB. Kadama, P. Nabyonga, J. Ogwal, PO. Nabukhonzo, P. Aguilar, AM (2006). 'Understanding the impact of eliminating user fees: utilization and catastrophic health expenditures in Uganda.' *Social Science & Medicine*, 62(4): 866–876.

Xu, K. Evans, DB. Carrin, G. Aguilar-Rivera, AM. Musgrove, P. Evans, T (2007). 'Protecting households from catastrophic health spending.' *Health Affairs (Project Hope)*, 26(4): 972–983.

2.5 *Health systems responsiveness: a measure of the acceptability of health-care processes and systems from the user's perspective*

NICOLE VALENTINE, AMIT PRASAD, NIGEL RICE, SILVANA ROBONE, SOMNATH CHATTERJI

Introduction

The World Health Organization (WHO) developed and proposed the concept of responsiveness, defining it as aspects of the way individuals are treated and the environment in which they are treated during health system interactions (Valentine et al. 2003). The concept covers a set of non-clinical and non-financial dimensions of quality of care that reflect respect for human dignity and interpersonal aspects of the care process, which Donabedian (1980) describes as "the vehicle by which technical care is implemented and on which its success depends". Eight dimensions (or domains) are collectively described as goals for health-care processes and systems (along with the goals of higher average health and lower health inequalities; and non-impoverishment – as measured through other indicators): (i) dignity, (ii) autonomy, (iii) confidentiality, (iv) communication, (v) prompt attention, (vi) quality (of) basic amenities, (vii) access to social support networks during treatment (social support), and (viii) choice (of health-care providers).

Building on extensive previous work, this chapter directs the conceptual and methodological aspects of the responsiveness work in three new directions. First, the given and defined domains (Valentine et al. 2007) are used to link responsiveness (conceptually and empirically) to the increasingly important health system concepts of access to care and equity in access. The concept of equity used in this chapter was defined by a WHO working group with experts on human rights, ethics and equity. It is defined as the absence of avoidable or remediable differences among populations or groups defined socially, economi-

cally, demographically or geographically (WHO 2005). Health inequities involve more than inequality – whether in health determinants or outcomes, or in access to the resources needed to improve and maintain health. They also represent a failure to avoid or overcome such inequality which infringes human rights norms or is otherwise unfair. Second, it expands on the issue of measurement strategies. Third, the psychometric results of the responsiveness module from the WHS are compared with its survey instrument predecessor in the Multi-country Survey (MCS) Study.

The chapter concludes with analysis of the most recent results for responsiveness from the WHS for ambulatory and inpatient health-care services for sixty-five countries (with special reference to subsets of European countries) to see how European countries' health-care systems perform with respect to responsiveness.

Responsiveness operationalized as a population health concept

Responsiveness is measured using criteria related to the importance of users' views. Individuals who use (or decide not to use) the health-care system are viewed as the appropriate source of information on non-technical aspects of care. This approach implies measuring responsiveness through household or other types of user surveys rather than, for example, expert opinion or facility audits.

Concepts such as quality of life and general satisfaction are also measured in surveys. However, self-reports have the additional criterion that they should be linked to one or several actual experiences with health services in the respondent's recent past (previous year) and upon which they base their views. These experiences are usually based on some type of interaction with the health-care system including interaction with a specific person in that system; a communication campaign; or another type of health system event or action that did not entail direct personal interactions. This criterion places the focus on what actually happened during contact with the health-care system, rather than the respondent's satisfaction or expectations of the health-care system in general.

WHO (2000) broadly defines the health-care system as: 'all actions whose primary intent is to produce health'. The responsiveness measure proposed by WHO conceptually aims to measure the responsiveness of the *whole health-care system* to the *whole population* (Murray

& Frenk 2000). When the self-report measurement approach based on the criterion of an actual (recent) experience is combined with the concept of measuring the *whole* population's experience of the *whole* health-care system then the measurement challenges are multiplied. We outline aspects of these challenges below.

Spheres of health events

Seven different types of health events that require interactions with health-care systems or services are listed below. The list is intended to be relevant generically, regardless of the configuration of providers, financing, technology, medicines and human resources:

1. ambulatory care in response to acute needs;
2. ambulatory care for chronic conditions;
3. inpatient care for short-term stays (>24 hours; <3 months);
4. long-term institutionalized care e.g. for populations with mental illnesses, disabilities related to physical health conditions or elderly populations;
5. non-excludable public health interventions e.g. public health promotion for communities or population groups such as access to improved water and sanitation, smoking bans;
6. opportunities for participation in health system governance e.g. shaping the health system and issues affecting health;
7. administrative and financial transactions: e.g. ease of making payments for services and medicines or of obtaining medicines with prescriptions, receiving reimbursement from insurance if needed.

This list illustrates that the design of questions in household or user surveys and the actual survey coverage would require significant work to cover the entire typology of interactions and abide by the criterion of obtaining user reports. For example, individuals receiving long-term institutionalized care cannot respond to household surveys and require more targeted designs. Also, questions may need to be tailored to the specific institutional arrangements of services (including insurance coverage) for a particular country, region or sector.

Roles of the users

Given that the health-care system is a socially constructed system, individuals' interactions with that system will differ according to circumstances. These can be categorized into four non-mutually exclusive groupings. For any given time period, a single survey respondent may have experiences of interactions that relate to all, none or some of these roles:

a. a patient or user (with or without personal contact);
b. a patient or user *by proxy* e.g. chiefly for children, but also for people with mental illness or elderly persons;
c. a relative or close friend of a patient;
d. a member of society who uses health services but has not done so in the defined period of the previous year, and who has some ability to shape the structure of health institutions. This citizen role is facilitated by the mechanisms for social participation in decision-making on health.

Combining health events and user roles – interactions

The full range of interactions combines user roles and different types of health events. When these are stated explicitly they help policymakers to understand which aspects of responsiveness they are most interested in capturing. A strategy to measure all these combinations of interactions and user roles would need to identify the most important in order to avoid overburdening respondents. This breadth of responsiveness is operationally challenging and to date has not been undertaken systematically in any country. Nevertheless, from a heuristic point of view, it is important to observe the potential implications of a concept if operationalized fully. It is also vital to decide whether measurement is necessary for all domains of responsiveness or a more limited set. WHO designed the WHS responsiveness instrument to cover interactions represented by the combination of events and user roles matching the alphanumeric labels listed above - 1ab, 2ab, 3ab, and 6d (involvement in decision-making only).

Responsiveness and equity in access

The link between responsiveness and equity in access is important. It derives from the impact of service qualities described by the responsiveness domains on utilization patterns. An explicit framework that describes how responsiveness is linked to access to care via the care context and process can inform empirical work aimed at describing responsiveness across countries. Fig. 2.5.1 presents such a framework that builds on other frameworks in the literature covering the medical-care process (Donabedian 1973); access to care (Aday & Andersen 1974; Tanahashi 1978); utilization (Andersen 1995; Bradley et al. 2002); and the conceptual framework proposed to the Commission on Social Determinants of Health (Solar & Irwin 2007).

The framework has three broad components: (i) environment; (ii) agents defining need for care; and (iii) process of care and outcomes (Fig 2.5.1). The first two components delineate context and together define the need for care at the population level. Their development was informed by the Aday and Andersen framework (1974) of 'health policy'; 'population characteristics'; 'health service characteristics'; and 'utilization', with some adaptations. For example, the decision-making agents component in the Fig. 2.5.1 framework draws attention to the role of both providers and users in defining need and setting the context for utilization. It evokes three agency groupings: (i) providers and their accepted protocols (which may differ across countries); lay persons (with their socially accepted protocols/norms); and the specific epidemiological or biological agents which produce different responses from the other two groups of decision-makers.

Recognition of the separate groupings of providers and lay persons is an important innovation that was raised in the Solar and Irwin (2007) framework and the work of the Health Systems Knowledge Network of the Commission on Social Determinants of Health (Gilson et al. 2007). This distinction is important for understanding the context in which responsiveness is measured and the implications for policy discussions. Responsiveness reports on convenience of access or confidentiality will reflect different profiles of services which have been negotiated by decision-making groups in society. For example, midwives in one country may make home visits that are not part of population health needs in another. Differences are to be expected and

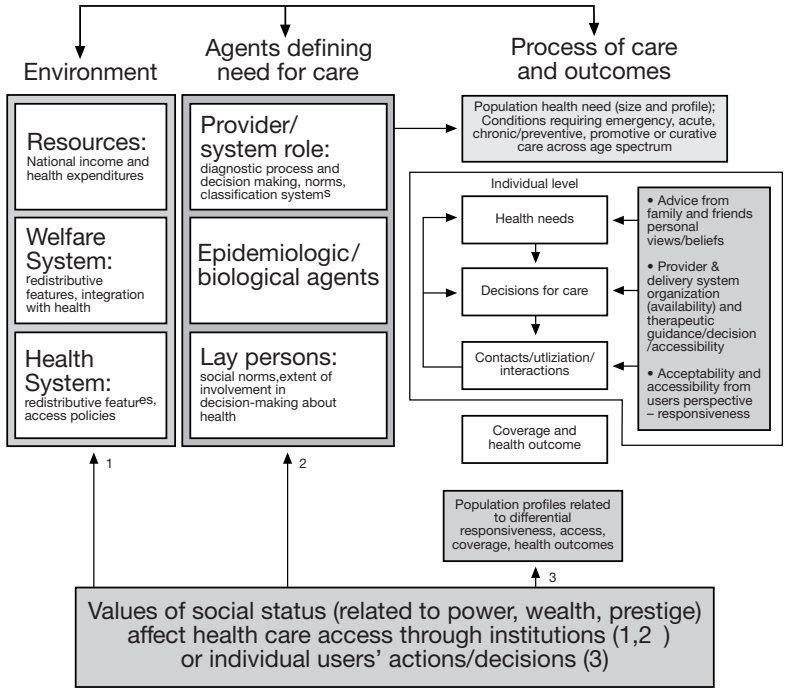

Fig. 2.5.1 Framework for understanding link between health system responsiveness and equity in access

may provide explanations for varying responsiveness across countries. However, it is important that these factors are explicit in analytical frameworks in order to understand how to improve responsiveness across different countries.

The third component of the framework is most relevant to the measurement of responsiveness – the process of care and outcomes. An individual who has a specific need for care moves from (a) recognition of health needs, to (b) decisions for care, to (c) contact with the system/utilization, and to (d) coverage. The latter is defined as the single, multiple or perpetual contacts to ensure adherence that may be required to guarantee adequate care for a particular condition (adapted from Tanahashi 1978). Care-seekers' decisions related to utilization and the possible achievement of full coverage (explained below) for a particular condition are influenced by three broad

factors shown in Fig. 2.5.1: (i) the *personal* context (advice from family and friends, personal beliefs); (ii) *providers* (administering therapeutic guidelines/decisions, organization of delivery e.g. being able to see a general doctor or specialist directly); and (iii) the health system's capacity to be responsive. The responsiveness domains mostly relate to Tanahashi's (1978) definitions of accessible (users able to reach and use health services) and acceptable care or coverage (users willing to use accessible services).

The concept of full coverage is introduced into the framework as coverage, although this term is used infrequently in the traditional access literature (except Tanahashi 1978 and, more recently, Shengelia et al. 2005). It usefully communicates the concept of a norm related to interventions for particular conditions. This differs from utilization rates for which high or low values indicate only the use of health-care resources without explicit reference to norms or need related to particular conditions. Health outcomes are affected by the extent of coverage reached and may not be affected by utilization rates. Of course, there is room for both concepts in the same framework as utilization rates for which the vulnerability of the population group is proxied (e.g. by income) do give some indication of the resources consumed relative to need.

The literature does make reference to definitions of coverage at population and individual levels. Shengalia et al. (2005) define *effective* coverage at the individual level as 'the fraction of maximum possible health gain an individual with a health care need can expect to receive from the health system.' Tanahashi (1978) refers to a population level measure of coverage as 'the number of people for whom the service has satisfied certain criteria relating to its intended health intervention, compared with the total target population.'

The third component of the framework also shows the links between responsiveness and equity in access. Responsiveness affects access at the individual level first. Responsiveness that is systematically worse for certain social groups with the same or greater need than other social groups could lead to inequities in access. These are defined as arising when anticipated, perceived or actual responsiveness attributes of the service dissuade certain social groups from seeking and receiving adequate care.[1] By adapting Tanahashi's (1978) population-level definition of coverage to the individual level, 'adequate care' would refer

[1] Definition suggested by Elias Mossialos, who commented on the draft chapter.

to services striving to meet a predefined technical norm in response to a variety of health conditions (completion of treatment; or continued, on-going treatment for chronic or palliative cases). Given this relationship between responsiveness, equity in responsiveness and equity in access, it is possible to use measures of responsiveness inequalities by different social groups (stratified according to need, e.g. proxied by income) to anticipate inequities in access.

Equity considerations for responsiveness survey design

A service that is perceived to have poor responsiveness may not be used optimally (or even at all) or as required by the health condition. Yet responsiveness measurement needs to be based on actual interactions. Thus, one weakness of the measurement approach is that measures will be biased upwards. This is not only because self-reports of this nature are usually biased upwards (see Ware & Hays 1988) but also because they do not fully capture the experiences of respondents who are in need but have not used services recently. Responsiveness measurement will not record the experience of care of someone who is excluded from care by failing to initiate (Aday & Andersen 1974) or obtain contact with the system (Tanahashi 1978).

Fig. 2.5.2 illustrates how populations may be excluded, with reference to two types of problems. In some cases, populations may not have sought care in the defined time period due to responsiveness or other factors e.g. financial barriers. These denied users would be excluded by screening questions on when they last came into contact with a health service. In other cases, the very nature of their vulnerability (e.g. homelessness) may put certain populations beyond the reach of traditional survey techniques. In both instances, surveys will be biased upwards and potentially underestimate inequalities in responsiveness. For the first problem, denied users can be asked about the barriers to care in order to gain qualitative information on the responsiveness measures. The second problem will require special survey efforts (e.g. surveys of institutionalized, homeless or migratory populations).

Special consideration should be given to the inclusion of service contacts with children as exposures at early stages of the life course have not only equity impacts that transmit into adulthood, but also intergenerational consequences. Minors cannot report for themselves but reporting by parents has been shown to be effective. This was used

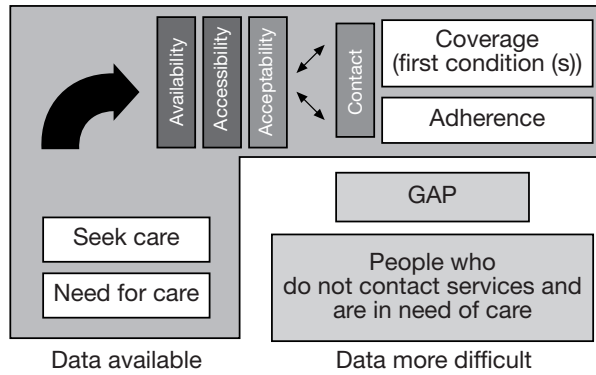

Fig. 2.5.2 Traditional survey methods omit data from certain population groups, overestimating responsiveness

Source: WHO & EQUINET (forthcoming)

for children up to the age of twelve in the WHS, as recommended by experts (WHO 2001).

Some critics have argued for special attention for sicker populations (Blendon et al. 2003) to ensure equity and because they know the services better. A strategy focusing on the sick may use health-facility exit-based surveys rather than household surveys, although this approach may omit those who have not used health services.

Responsiveness questionnaires

The responsiveness domains were derived from existing patient questionnaires and studies as reported in the extensive literature review conducted by De Silva (2000). This review profiled the questionnaire work undertaken by the AHRQ, Harvard Medical School, the Research Triangle Institute and the RAND Corporation. None of the existing questionnaires and studies captured all the dimensions that they covered collectively. WHO developed an instrument (questionnaire) that covered the collection of dimensions (described in the literature review) related to non-technical aspects of the process of care: dignity, autonomy, communication, confidentiality, prompt attention (related to convenience and peace of mind rather than urgent medical attention), quality of basic amenities, access to social support networks during treatment (labelled 'social support' in the MCS Study

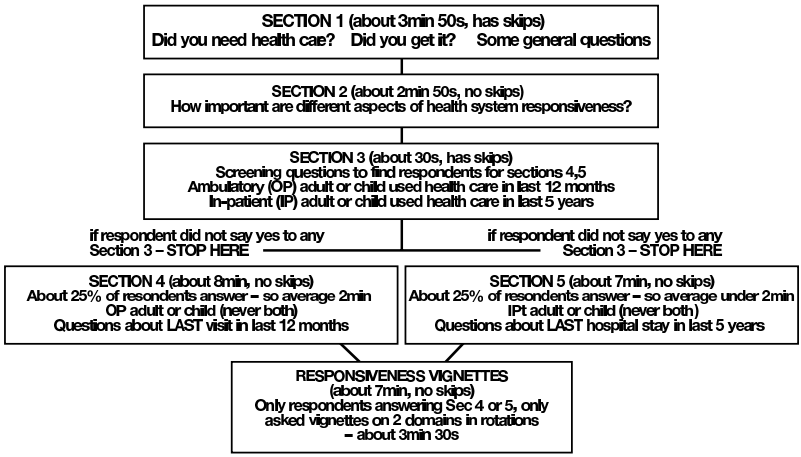

Fig. 2.5.3 Responsiveness questionnaire as a module in the WHS questionnaire: interview structure and timing

and 'access to family and community support' in the WHS) and choice (of health-care providers).

WHO's responsiveness questionnaire has been developed and refined. Questions (items) were initially fielded in a key informants' survey of thirty-five countries and the results described in *The world health report 2000* (WHO 2000). A household survey instrument which included pre-testing was then developed as part of the MCS Study covering sixty countries (Ustun et al. 2001; Valentine et al. 2007). Following the launch of the MCS Study, the concept of responsiveness and the questionnaire were refined and a revised instrument was included in the WHS implemented across seventy countries in 2002–2003.

The WHS basic survey mode used an in-person interview conducted in one of three possible forms: ninety-minute in-household interview (fifty-three countries) (long-form); thirty-minute face-to-face interview (short-form) (thirteen countries); or computer-assisted telephone interview. Samples were randomly selected (those above eighteen years) resulting in sizes of between 600 and 10 000 for each country surveyed. Descriptive statistics about individuals sampled in each country are reported in Annex 2. Data collection was performed on a modular

basis, addressing different aspects of health and the health system and including information on health insurance, health expenditures, socio-demographics and income, health state valuations, health system responsiveness and health system goals (Üstün et al. 2003). Fig. 2.5.3 provides an overview of the responsiveness module in the WHS. The measurement of responsiveness was obtained by asking respondents to rate their most recent experience of contact with the health system within each of the eight domains by responding to the set of questions listed in Fig. 2.5.4. The response categories available were very good, good, moderate, bad and very bad.

Like health, responsiveness is viewed as a multidimensional concept. Each domain is measured as a categorical variable for which there is an assumed underlying latent scale. Certain domains are more suited to patient evaluation, e.g. quality of basic amenities and prompt attention. In contrast, it is more difficult to evaluate whether full details of the nature of an illness and all relevant treatments and available options have been disclosed as this requires specialist knowledge. Accordingly, it is more problematic to maintain objectivity in the evaluation of some domains. Samples have undergone extensive quality assurance pro-cedures at data collection stage at country and inter-country levels.

The MCS Study and WHS modules on responsiveness have strong similarities. However, they have a number of different ways of expanding coverage and alleviating the burden on survey respondents. More notable changes in the WHS include: more face-to-face interviews or computer-assisted telephone interviews (MCS Study included twenty-eight postal surveys); eliciting the experiences of children up to twelve (reported through a parent); and reducing the number of items that individuals are required to respond to on each domain. The WHS module also tried to identify barriers to access by asking people if they needed care and, if so, whether they sought care or why they did not (Fig. 2.5.3 section 1). The analyses that follow focus on the questions asked in sections four and five of the responsiveness module and cover the ambulatory and hospital (inpatient) experiences of adult and child populations.

Fig. 2.5.4 Operationalization of responsiveness domains in the WHS

Responsiveness domain label (short description)	**Item questions**
Prompt attention (convenient travel and short waiting times)	How would you rate: 1- travelling time to the hospital 2- time you waited before being attended to[b]
Dignity (respectful treatment and communication)	How would you rate: 1- being greeted and talked to respectfully[a] 2- respect for privacy during physical examinations and treatments[a,b]
Communication (clarity of communication)	How would you rate: 1- how clearly health-care providers explained things to you[a] 2- the time you get to ask questions about your health problems or treatment[a,b]
Autonomy (involvement in decisions)	How would you rate: 1- being involved in making decisions about your health care or treatment[a] 2- the information you get about other types of treatments or tests[b]
Confidentiality (confidentiality of personal information)	How would you rate the way: 1- health services ensured you could talk privately to health-care providers[a] 2- your personal information was kept confidential[a,b]
Choice (choice of health-care provider)	How would you rate: 1- your freedom to choose the health-care providers that attended to you
Quality of basic amenities (surroundings)	How would you rate: 1- cleanliness of the rooms inside the facility, including toilets[a] 2- amount of space you had[a,b]
Access to family and community support (contact with outside world and maintenance of regular activities)	How would you rate: 1- ease of having family and friends visit you 2- experience of staying in contact with the outside world when you were in hospital[a,b]

[a] Similar items appear in the MCS Study
[b] Item omitted from short version of WHS

Psychometric properties of the responsiveness domain questions

Psychometrics examines the quality of survey instruments and has been used extensively to assess the quality of the responsiveness instrument in both the MCS Study and the WHS. This section briefly considers three key desirable properties of a survey instrument (feasibility, reliability, validity) and compares them in the MCS Study and the WHS. The results on these properties are presented in combination for ambulatory and home care (as ambulatory care) and separately for inpatients. A more detailed description of the psychometric properties of the MCS Study is provided by Valentine et al. (2003a & 2007).

Feasibility

Feasibility refers to the ease of administering an instrument in the field and can be assessed by considering factors such as survey response rates, the proportion of missing items in a respondent interview (inappropriate missing responses) and item missing rates (percentage of respondents who omitted a particular item). The literature provides little indication of an acceptable survey response or inappropriate response missing rates but, in general, guidance indicates that item missing rates below 20% can be considered acceptable (Valentine et al. 2007; WHO 2005a).

Survey response rates measured as a percentage of attempted and effective contacts were available only for the MCS. The comparison of reliability between the two surveys rests mainly on interview completion (a form of survey response rates) and item missing rates. It is important to note that interview completion rates may be as high as 100% as they give the number of persons who started and completed interviews as a percentage of the number of persons who started interviews.

The MCS Study shows high measures of feasibility with a response completion rate greater than 95% for each of the countries considered, except Colombia (73%). Furthermore, no country exceeded a 20% item missing rate and only three countries had item missing rates in excess of 10% (Switzerland, Turkey, Tobago). Valentine et al. (2007) provide full results of the psychometric properties of the MCS

Table 2.5.1 ***Item missing rates, ambulatory care (%)***

	MCS Study	WHS
Prompt attention	0.86	1.72
Dignity	1.13	1.75
Communication	0.55	0.38
Autonomy	2.70	2.03
Confidentiality	6.40	2.43
Choice	7.50	3.25
Quality of basic amenities	2.30	3.25
Average	3.06	2.12

Study. A similar analysis of the responsiveness instrument in the WHS showed that response completion rates per country were greater than 80% for all countries except Israel (63%). No country exceeded the accepted item missing rate threshold of 20% for ambulatory care and only Swaziland exceeded this threshold for inpatient care.

Additional information on the feasibility of the WHS responsiveness instrument is provided by the percentage of respondents that report missing values for zero; one; two; or three or more items. In countries where the long-form questionnaire was implemented, in responses on ambulatory care 88% of respondents reported no missing items; 6% reported one; 2% reported two and 4% reported three or more. Corresponding values for inpatient care were 87%, 5%, 1% and 7%. In countries where the short-form questionnaire was implemented, in responses for ambulatory care 87% returned no missing items, 11% reported one, 3% reported two and 2% reported three or more. The corresponding figures for inpatient care are 81%, 11%, 4% and 4%.

Table 2.5.1 offers a more direct comparison of the item missing rates. The values for the MCS Study are taken from Valentine et al. (2007) and consider only the forty-one countries in which interviewer administered interviews were held, corresponding to the method used in the WHS. Item missing rates are provided for ambulatory care by domain (calculated as the arithmetic mean of missing rates of individual items present in a domain) by averaging across countries. As can be seen, the WHS reported lower item missing rates for four

of the seven domains and failed to exceed 3.25% in any domain. Averaged across countries and domains, the overall missing item rate in the WHS is nearly 1% lower than that in the MCS Study.

Reliability

The reliability of an instrument refers to the test-retest property of measurement, usually over time, all other things being equal. Temporal reliability can be measured using the kappa statistic. Landis and Koch (1977) suggest that statistics in the range 0.41–0.60 indicate moderate reproducibility; 0.61–0.80 substantial reproducibility and 0.81–1.00 almost perfect reproducibility.

Instrument reliability in the MCS Study was assessed by re-administering the entire responsiveness questionnaire to respondents in ten country sites one month after the initial interview. There is high reliability of all items by domain when averaged across the countries (see Valentine et al. 2007). The lowest kappa value reported for any domain was 0.64 (for dignity in home care). However, there is variability in reliability when results are averaged across domains within countries. Reproducibility is substantial in five countries, moderate in three and low in two.

The reliability of the WHS instrument was assessed by re-interviewing 10% of the original sample in each country. The re-interviewed respondents were selected randomly and asked to complete the follow-up questionnaire one to seven days after the first interview (Üstün et al. 2005). We consider reliability in fifty-three countries for ambulatory care and fifty-five countries for inpatient care where sufficient data points (>20) were available in the follow-up survey. When the kappa statistics are averaged across items within countries, at least moderate reliability was reported for ambulatory care in twenty-four countries and for inpatient care in twenty-seven countries. When results are averaged across countries for each item separately all items satisfy at least the condition for moderate reproducibility.

Table 2.5.2 compares kappa statistics for the MCS Study and the WHS. The kappa statistic is provided for each domain, averaged across countries and overall for countries and domains. The first and second columns in Table 2.5.2 show kappa statistics averaged across the ten countries in the MCS Study and the fifty-three countries of the WHS in which the responsiveness instrument was re-administered to respon-

Table 2.5.2 ***Reliability in MCS Study and WHS***

	MCS+ (10 countries)	WHS (53 countries)	MCS+ (India, China)	WHS (India, China)
Prompt attention	0.60	0.49	0.66	0.73
Dignity	0.61	0.45	0.69	0.71
Communication	0.57	0.45	0.67	0.73
Autonomy	0.65	0.46	0.71	0.70
Confidentiality	0.59	0.45	0.74	0.71
Choice	0.63	0.40	0.75	0.72
Quality of basic amenities	0.65	0.44	0.71	0.72

+*Source*: Valentine et al. 2007

dents. When considering all available countries, the kappa statistics are considerably lower for the WHS. However, this does not provide a like-for-like comparison. Consideration of the two countries common to both surveys (India and China) provided in columns three and four indicates very similar comparisons of reliability in each survey.

Psychometric measures can also be investigated where data are stratified by population groups of interest. This allows an assessment of whether any revealed systematic variations suggest caution in interpreting results or indicate a need for greater testing before a survey is implemented.

We investigated the reliability of the WHS responsiveness instrument across European countries for two population groups defined by educational tenure. Table 2.5.3 presents average kappa statistics for each domain separately for western European countries and those of Central and Eastern Europe and the former Soviet Union (CEE/FSU) (listed in Annex 1). Results are further presented by level of educational tenure (defined as people having studied for either more or less than twelve years). Table 2.5.3a and Table 2.5.3b report results for ambulatory care and inpatient care, respectively. Overall, the reliability of the responsiveness instrument appears to be greater in CEE/FSU countries than in western European countries, irrespective of levels of education.

Table 2.5.3a ***Reliability across European countries: ambulatory care***

	Western Europe		CEE/FSU		Europe overall	
	Education		Education		Education	
	Low	High	Low	High	Low	High
Prompt attention	0.49	0.44	0.59	0.56	0.54	0.50
Dignity	0.40	0.40	0.57	0.60	0.49	0.50
Communication	0.42	0.42	0.52	0.49	0.47	0.45
Autonomy	0.43	0.41	0.55	0.46	0.49	0.43
Confidentiality	0.25	0.52	0.58	0.52	0.41	0.52
Choice	0.37	0.26	0.61	0.52	0.49	0.39
Quality of basic amenities	0.24	0.37	0.54	0.53	0.39	0.45
Average	0.37	0.40	0.56	0.52	0.47	0.46

Table 2.5.3b ***Reliability across European countries: inpatient care***

	Western Europe		CEE/FSU		Europe overall	
	Education		Education		Education	
	Low	High	Low	High	Low	High
Prompt attention	0.30	0.38	0.68	0.53	0.49	0.45
Dignity	0.34	0.40	0.65	0.53	0.50	0.47
Communication	0.25	0.34	0.56	0.52	0.41	0.43
Autonomy	0.19	0.24	0.61	0.48	0.40	0.36
Confidentiality	0.21	0.37	0.60	0.49	0.41	0.43
Choice	0.23	0.34	0.64	0.49	0.43	0.42
Quality of basic amenities	0.29	0.43	0.62	0.52	0.46	0.47
Social support	0.26	0.38	0.60	0.49	0.43	0.43
Average	0.26	0.36	0.62	0.51	0.44	0.43

CEE: Central and eastern Europe; FSU: Former Soviet Union

Interestingly, country groupings indicate that the reliability of the instrument is greater for less educated individuals in CEE/FSU countries but generally the opposite appears to hold for western Europe. Taken in their totality across both groups of countries, the results suggest that (with the exception of the domain for confidentiality and choice) educational achievement has little influence on the reliability of the responsiveness instrument. Further, the reliability of the instrument for ambulatory care appears marginally better than for inpatient care (except for quality of basic amenities domain).

Validity

The psychometric property of validity focuses on exploring the internal structure of the responsiveness concept, particularly the homogeneity or uni-dimensionality of responsiveness domains. The property is often measured through factor analysis and Cronbach's alpha. Stronger evidence of uni-dimensionality (factor loadings close to +1 or -1) supports greater validity of the instrument; a minimum value in the range of 0.6 to 0.7 has been suggested for Cronbach's alpha (e.g. Labarere 2001; Steine et al. 2001).

Validity was assessed by pooling data from different countries and analysing each domain independently. For the MCS Study, values of Cronbach's alpha suggested that all domains lay within the desired range and were greater than 0.7 for all except one (prompt attention = 0.61) (Valentine et al. 2007). For the WHS all countries satisfied the requirement that Cronbach's alpha is greater than 0.6 – the minimum value across countries was 0.66 for inpatient care and 0.65 for ambulatory care. This requirement was also satisfied for all domains except prompt attention for ambulatory care (alpha = 0.56).

We further evaluated the construct validity of the WHS questionnaire using maximum likelihood exploratory factor analysis, as performed by Valentine et al. (2007) when analysing the MCS Study ambulatory responsiveness questions (inpatient sector of MCS Study contained only one item per domain, except for prompt attention and social support). The method makes reference to Kaiser's eigenvalue rule which stipulates that item loadings on factors should be 0.40 or greater (Nunnally & Bernstein 1994). The results of the MCS Study analysis are presented by Valentine et al. (2007).

Table 2.5.4 *Promax rotated factor solution for ambulatory responsiveness questions in the WHS*

Domain	Item	Latent underlying factor							Uniqueness
		1	2	3	4	5	6	7	
Prompt attention	1	-0.018	0.115	0.135	-0.006	0.056	-0.013	0.288	0.774
	2	0.010	-0.019	-0.038	0.013	-0.019	0.019	**1.023**	0.000
Dignity	1	0.048	**0.728**	0.045	-0.046	0.044	-0.027	0.061	0.352
	2	0.025	**0.719**	-0.079	0.225	-0.009	-0.003	-0.041	0.311
Communication	1	0.523	0.321	-0.063	0.076	0.048	0.014	-0.014	0.327
	2	**0.855**	-0.017	0.038	0.000	0.048	0.011	0.019	0.157
Autonomy	1	**0.476**	-0.027	0.042	0.020	0.371	0.034	0.021	0.294
	2	-0.011	0.029	0.010	-0.005	**0.924**	0.028	-0.017	0.116
Confidentiality	1	0.072	0.039	-0.030	**0.614**	0.194	0.013	0.032	0.327
	2	-0.005	0.028	0.033	**0.849**	-0.050	-0.005	0.010	0.257
Choice	1	0.072	-0.055	**0.629**	0.037	0.145	-0.042	-0.038	0.462
Facilities	1	-0.021	0.169	0.185	0.134	-0.058	**0.444**	-0.043	0.462
	2	0.016	-0.050	-0.063	-0.028	0.034	**1.052**	0.026	0.000

Valentine et al's (2007) results confirmed the hypothesized domain taxonomy for the majority of the domains. The high human development countries have a few exceptions within the domains of prompt attention and dignity, where items tend to load on multiple factors. For the WHS questionnaire, Table 2.5.4 reports the promax rotated factor solutions for ambulatory care computed across all countries (pooled) in which the long-form questionnaire was implemented.[2] In general, results confirmed the hypothesized domain taxonomy, as the items belonging to particular domains (except autonomy) loaded on a single factor. For autonomy, the largest loading for the first item was on the factor for communication but the second largest loading (0.371) corresponded to the largest loading on the second item (factor 5). For prompt attention, the two largest loadings fell on a single factor (7) but did not reach the threshold suggested by Nunnally and Bernstein (1994).

As seen in Table 2.5.5, the hypothesized domain taxonomy was also confirmed for inpatient care and, again, the items failed to load on a single factor in only two domains (prompt attention, communication). The communication item related to information exchange loaded more strongly on the autonomy domain. In general, the strong association between autonomy, communication and dignity domain items supports the assertions made in previous MCS Study work and elsewhere that communication is an important precondition or accompaniment to being treated with dignity and involvement in decision-making about care or treatment.

Measuring responsiveness

Calculating the measures

Two measures are used to capture health system responsiveness in the analyses that follow. The first is the level of responsiveness; the second is the extent of inequalities in responsiveness across socio-economic groups in a country. This second measure can be used as a proxy for equity in responsiveness as explained below. Both measures are applied to user reports from ambulatory and inpatient health-care settings, resulting in four indicators per country.

[2] This type of analysis is not suitable for countries in which the short-version questionnaire was implemented as only one item was present in each domain.

Table 2.5.5 *Promax rotated factor solution for inpatient responsiveness questions in the WHS*

		Latent underlying factor										
Domain	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Uniqueness**
Prompt attention	1	0.009	0.002	-0.073	-0.011	0.005	-0.004	-0.011	-0.011	**1.041**	0.007	0.000
	2	-0.007	-0.004	**0.446**	0.063	-0.021	0.031	0.051	0.044	0.233	-0.037	0.543
Dignity	1	0.036	-0.051	**1.007**	-0.023	-0.018	-0.007	0.014	-0.012	-0.081	0.005	0.134
	2	0.052	0.263	**0.437**	0.008	0.172	0.024	-0.099	-0.002	0.010	0.029	0.371
Communication	1	0.150	-0.016	0.038	0.004	**0.786**	0.005	0.019	0.022	0.009	-0.005	0.131
	2	**0.526**	-0.002	0.032	0.003	0.144	0.015	0.025	0.292	-0.012	-0.001	0.239
Autonomy	1	**0.757**	0.040	-0.009	0.028	-0.021	0.009	-0.030	0.167	-0.002	0.002	0.253
	2	**0.951**	-0.011	0.046	-0.004	0.009	0.010	0.017	-0.219	0.028	-0.004	0.184
Confidentiality	1	0.178	**0.632**	0.011	0.098	0.032	0.010	0.028	-0.022	-0.034	-0.134	0.307
	2	0.026	**0.874**	-0.016	-0.055	-0.033	0.017	0.009	0.014	0.021	0.013	0.269
Choice	1	0.254	0.053	0.006	0.007	0.021	0.024	**0.475**	0.007	-0.017	0.012	0.455
Facilities	1	0.026	0.091	0.060	**0.501**	0.004	0.034	0.067	-0.013	0.019	0.141	0.417
	2	0.017	-0.045	-0.037	**0.959**	-0.002	0.035	-0.032	0.007	-0.014	0.007	0.147
Social support	1	-0.014	0.031	0.029	0.121	0.016	**0.747**	-0.027	-0.019	-0.011	-0.003	0.294
	2	0.039	-0.011	-0.021	-0.034	-0.010	**0.871**	0.024	0.016	0.006	0.003	0.244

The level of responsiveness (also called the responsiveness score) is calculated by averaging the percentage of respondents reporting that their last interaction with the health-care system was good or very good across the relevant domains (seven domains for ambulatory care; eight for inpatient). This average is referred to as overall ambulatory or inpatient responsiveness. A higher value indicates better responsiveness. Scores or rates per country are age-standardized using the WHO World Standard Population table, given that increasing age is associated with increasingly positive reports of experiences with health services (Hall et al. 1990).

The inequality measure is based on the difference across socio-economic groups, in this case identified by income quintiles and a reference group.[3] From a theoretical perspective, the reference group could be chosen on the basis of the best rate in the population; the rate in the highest socio-economic group; a target external rate; or the mean rate of the population. The highest income quintile reference group was selected here. Each difference between the highest and other quintiles is weighted by the size of the group with respect to the reference group. The measure is calculated for each domain and an average is taken across all domains to derive a country inequality indicator (again, for ambulatory or inpatient services separately).[4] Higher value for the inequality measure indicates higher inequalities and, by proxy, higher inequities (see below).

The assumption behind the link between the inequality measure of responsiveness calculated here and an inequity measure is based on the equity criterion that there should be an equal level of responsiveness for people with equal levels of health need. To the extent to which income may proxy as health needs (assuming a negative relationship between income and ill-health), then a positive gradient between income quintiles and responsiveness levels provides evidence of inequity. In other

[3] Harper, S. Lynch, J (2006). Measuring health inequalities. In: Oakes, JM. Kaufman, JS (eds.). *Methods in social epidemiology*. San Francisco: John Wiley & Sons. The indicator was further modified by Dr. Ahmad Hosseinpoor (WHO/IER). The title of the paper is "Global inequalities in life expectancy among men and women" (tentative).

[4] The formula: $\sum_{j=1}^{J} N_j \left| y_j - \mu \right| \Big/ N$; y_j *: the rate in group j,*μ *: the rate in reference group,* N_j *: population size of each group,N: Total population*

words, a positive gradient from low to high income groups would imply inequities in responsiveness. Lower income groups would presumably have greater health service needs and be entitled to at least the same, or better, responsiveness from the health system.

All domain results were sample weighted and average responsiveness scores were age-standardized because of the widespread evidence of a systematic upward bias in rating in the literature and reports on responsiveness and quality of care in older populations (Valentine et al. 2007).

Interpreting the measures

In interpreting the indicators of responsiveness, there is no clear cut-off between acceptable and unacceptable. Clearly, higher responsiveness levels and lower inequality measures are better. The literature shows that self-reported measures (e.g. responsiveness, quality of life, satisfaction) are right-skewed. This was illustrated in the WHO's raw survey results in which 81% of respondents reported in the highest two categories (range 52%-96%) in the MCS Study and an average of 72% (range 38%-92%) in the WHS. Therefore, the framework for interpreting the results on the WHS presented here adopts a benchmarking approach, comparing countries with similar resource levels based on the World Bank income classification of countries (see Annex 1, Fig. A). The WHS classification of countries was incorporated for the European results – western European, and eastern European and former Soviet Union countries (Annex 1, Fig.B).

Using this benchmarking approach and the analytical framework shown in Fig. 2.5.1, we had some expectations of how the WHS results would look. We expected responsiveness to be greater in high resource settings because of the increased availability of human resources and better infrastructure. Human resources are the main conduit for the respect of person domains and, to some degree, prompt attention and choice. The higher the quality of the basic infrastructure in a country (e.g. better transport networks) the greater the impact on the domains of prompt attention and quality of basic amenities in health services.

We anticipate that there will be differences between responsiveness measures and general satisfaction measures for the same country although no direct comparison is drawn in this chapter. Measures of general satisfaction may respond to the contextual components

described in Fig. 2.5.1 but measures of responsiveness are based on actual experiences and will reflect the care process from the perspective of users.

WHS 2002 results

Sample statistics

The WHS 2002 was conducted in seventy countries, sixty-nine of which reported back to WHO on their responsiveness data. Turkey did not complete the responsiveness section. The average interview completion response rate was 91% for all countries, ranging from 44% for Slovenia and up to 100% for as many as twenty-two countries. Note that the measure of survey response rates was interview completion rates – as mentioned, these may be as high as 100% as they express the number of persons who started and completed interviews as a percentage of the number of persons starting interviews. Sample sizes for ambulatory and inpatient care services averaged 1530 and 609 respectively, across all countries. A wide range across countries (130–19 547 for ambulatory use in the last twelve months; 72–1735 for inpatient use in the last three years) depended on both overall survey samples and different utilization rates across the different countries. Female participation in the overall survey sample averaged 56%, ranging from 41% (Spain) to 67% (Netherlands). The average age across all surveys was forty-three, ranging from thirty-six in Burkina Faso to fifty-three in Finland. Details on country-specific samples are provided in Annex 2.

Ambulatory care responsiveness

All countries

Overall results followed expected trends,[5] with higher overall levels of responsiveness in higher-income countries as shown in Fig. 2.5.5. Inequalities between lower- and middle-income countries changed slightly but, in general, large reductions in inequalities were only observed when moving from middle- to high-income countries.

[5] Australia, France, Norway and Swaziland were not included as they did not record an ambulatory section. Italy, Luxembourg, Mali and Senegal were dropped as their datasets lacked (minimum) sufficient observations for each quintile (thirty or more).

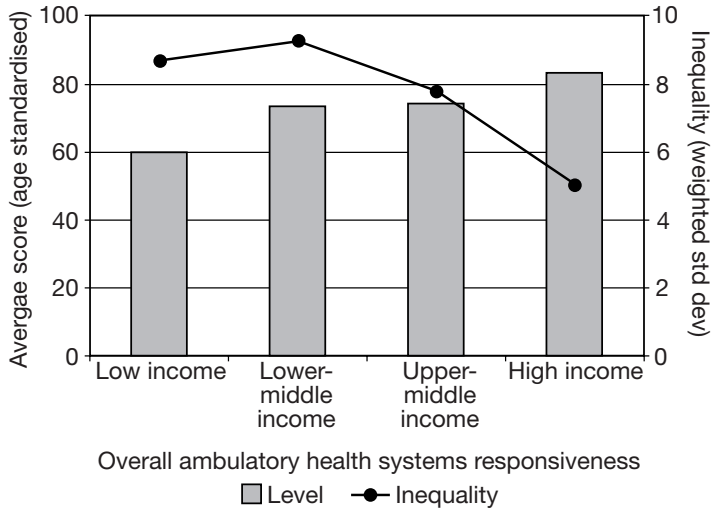

Fig. 2.5.5 Level of inequalities in responsiveness by countries grouped according to World Bank income categories

Respondents from different country groupings consistently reported low responsiveness levels and high inequalities for the prompt attention domain. The dignity domain was consistently reported as high and with low inequalities. The overall gradient between country groupings as described in Fig. 2.5.5 held for all domains. In other words, no domain was performing significantly better in a lower income grouping of countries than in the higher income grouping.

European countries

Within Europe, western European countries showed notably higher mean levels of responsiveness and lower inequalities than the CEE/FSU countries (Fig. 2.5.6). Responsiveness levels across all twenty-five European countries ranged from 56% in Russia to 92% in Austria (Fig. 2.5.7). Inequalities ranged from 2.2 in Spain to 14.3 in Bosnia and Herzegovina. Strikingly, nine of the twelve CEE/FSU countries had inequalities higher than the European average and only four of the twelve CEE/FSU countries had responsiveness levels greater than the average levels for Europe as a whole. By contrast, twelve of the thirteen western European countries had responsiveness levels higher than the European average.

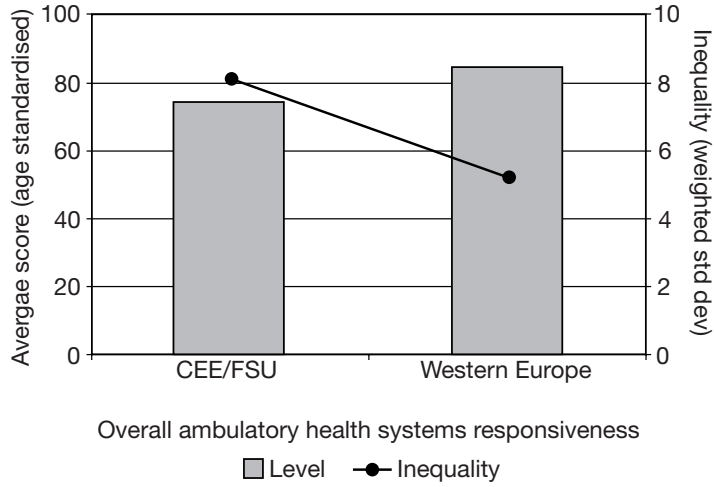

Fig. 2.5.6 Level of inequalities in responsiveness by two groups of twenty-five European countries

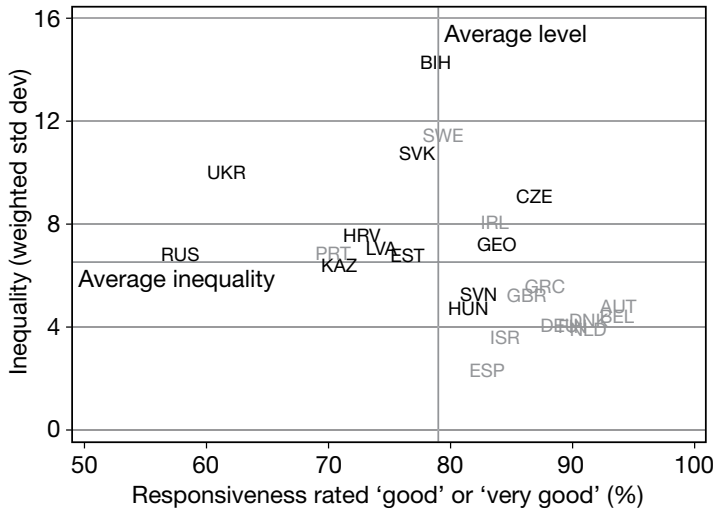

Fig. 2.5.7 Inequalities in ambulatory responsiveness against levels for twenty-five European countries

On average, responsiveness for all domains in western European countries was higher than in CEE/FSU countries. Differences were largest for the choice and autonomy domains. Prompt attention was the worst performing domain in western Europe, while autonomy and prompt attention were the worst performing domains in CEE/FSU countries. Dignity was the best performing domain in both groups of countries, as found for the global average.

Inequalities were higher for all domains in CEE/FSU countries. Both groups of countries had the highest inequalities in the prompt attention domain. Inequalities were lowest in the communication domain in CEE/FSU countries and in the basic amenities and dignity domains in western Europe.

Inpatient health services

All countries

The level of responsiveness for inpatient services increased across the four income groupings of countries (Fig. 2.5.8).[6] However, the pattern for inequalities was surprising. Unlike the trend in ambulatory care, inpatient inequalities reached a peak in upper middle-income countries (greatest values in South Africa and Slovakia).

Responsiveness domain levels (except for autonomy and choice) increased across country groupings. Upper middle-income countries had lower levels of both domains than lower middle-income countries. In general, these domains were also the worst performing (compared with prompt attention for ambulatory services). The dignity domain performed best in all groupings of countries, followed closely by social support. The spike in inequalities observed for upper middle-income countries seems to have arisen from sharply higher inequalities for the autonomy, basic amenities and social support domains.

European countries

For ambulatory services, responsiveness levels and inequalities in inpatient services differed between western Europe and CEE/FSU countries

[6] Australia, France and Norway were not included because they lacked data on assets necessary for construction of wealth index; Swaziland had too few observations in the ambulatory section. Ethiopia, Italy, Mali, Senegal and Slovenia were dropped from the analysis as their datasets did not have (minimum) sufficient observations for each quintile.

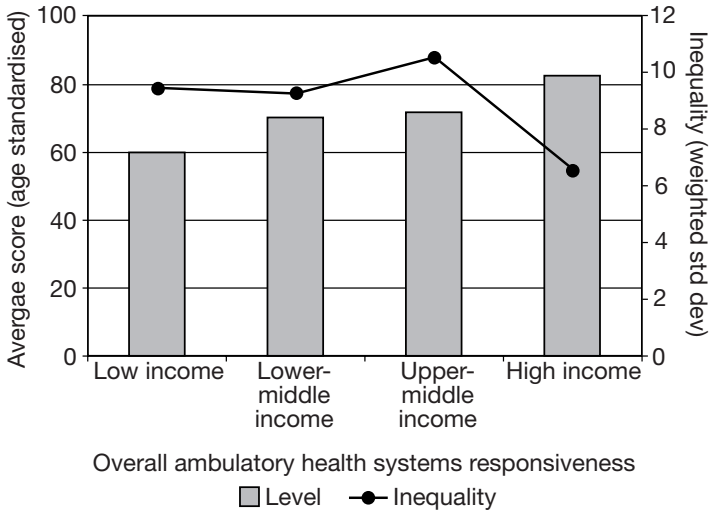

Fig. 2.5.8 Level of inequality in responsiveness across World Bank income categories of countries

(Fig. 2.5.9). The average level of responsiveness levels across eleven CEE/FSU countries is 70% compared to 80% for fourteen countries in western Europe.[7] Inequalities were also higher in CEE/FSU countries.

Across all twenty-five European countries, responsiveness levels range from 51% in Ukraine to 90% in Luxembourg. Inequities range from a low of 3.4 in Austria to 18.9 in Slovakia. Ten of the eleven CEE/FSU countries (shown in grey in Fig. 2.5.10) have responsiveness inequalities higher than the European average (for inequalities). Only five of the eleven CEE/ FSU countries have responsiveness levels higher than the average level for Europe, whereas all fourteen western European countries have a responsiveness level higher than the European average.

As for ambulatory services, western European countries show higher levels for each of the eight domains of inpatient services. Dignity was the best performing domain in CEE/FSU countries; in western Europe both dignity and social support had the highest (similar) levels. Choice was the worst performing domain for both groups of countries.

[7] Italy and Slovenia were omitted from the inpatient services analysis as their datasets did not have the minimum number of observations required for reliable results.

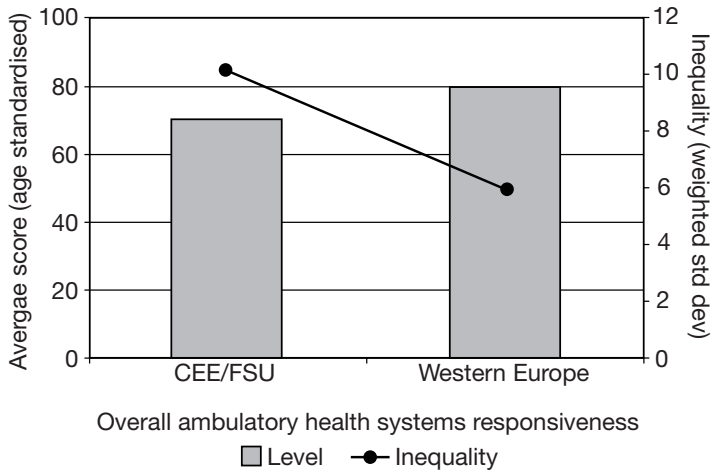

Fig. 2.5.9 Level of inequalities in responsiveness by two groups of twenty-five European countries

Inequalities in all domains were higher for CEE/FSU countries; the highest inequality was seen in the prompt attention domain. In western Europe, inequalities were highest in the domains of autonomy and confidentiality. In CEE/FSU countries the lowest inequalities were seen in the dignity domain while in western Europe the lowest inequalities were seen in social support.

Responsiveness gradients within countries

Ambulatory health services

The values for the inequality indicator ranged between five and ten for the different groups of countries. Fig. 2.5.11 shows how these values translate into a gradient in responsiveness for different wealth or income quintiles within countries. Low- and middle-income countries showed a gradient but no gradient was seen in the high-income countries when averaged together.

In Europe, the CEE/FSU countries showed a gradient in the level of responsiveness across wealth quintiles with richer populations reporting better responsiveness (Fig. 2.5.12). The gradient was nearly flat for western European countries.

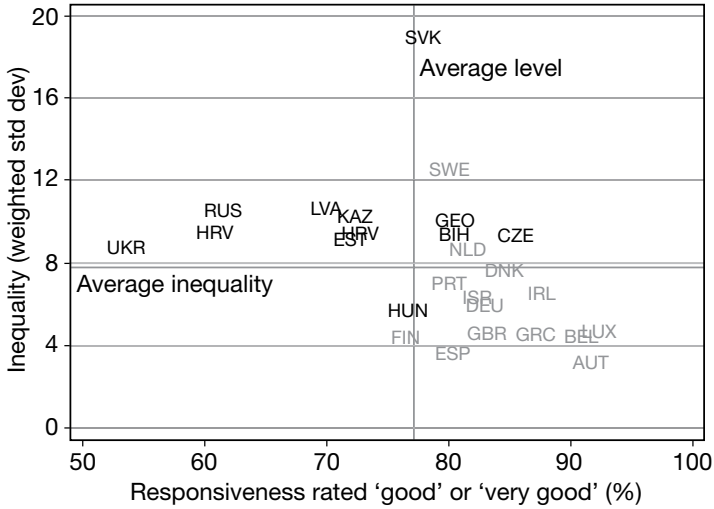

Fig. 2.5.10 Responsiveness inequalities against levels for twenty-five European countries

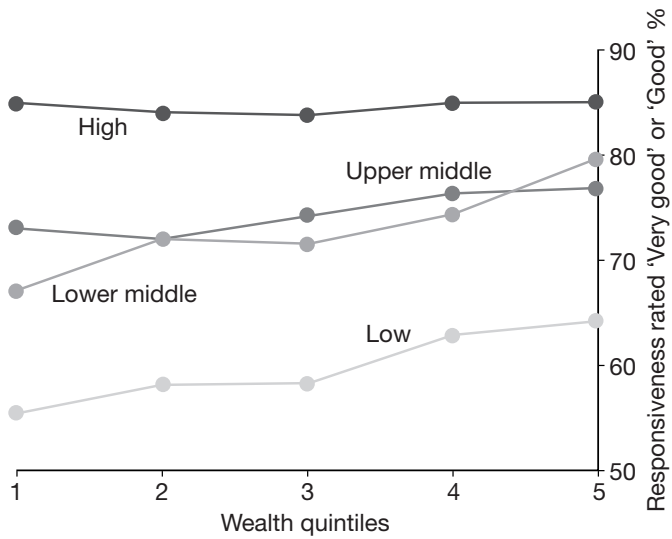

Fig. 2.5.11 Gradient in responsiveness for population groups within countries by wealth quintiles

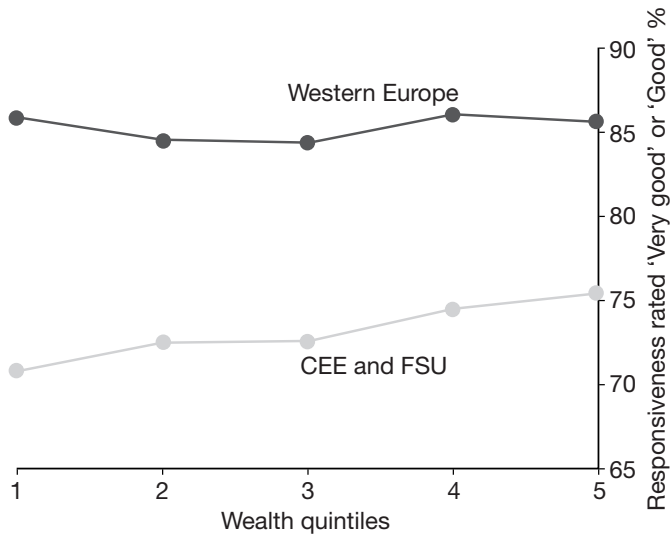

Fig. 2.5.12 Gradient in responsiveness for population groups within countries in Europe by wealth quintiles

Inpatient health services

The gradient in responsiveness for inpatient services is flatter than that observed for ambulatory services and most marked in low-income countries (Fig. 2.5.13). Similarly, no gradient can be observed across wealth quintiles in the two groups of European countries. However, people in all quintiles in CEE/FSU countries clearly face worse levels of responsiveness than people in any quintile of western Europe (Fig. 2.5.14).

Health system characteristics and responsiveness

Fig. 2.5.1 shows the rather obvious observation that factors such as resources in the health system provide a context to the process of care. It also shows the less obvious result that responsiveness affects the process of care, especially with respect to completion of treatment. We refer to this as coverage. With this understanding, we first explored the relationship between health expenditure and responsiveness in order to assess which domains might be more affected. Second, we explored the relationship between responsiveness and indicators of

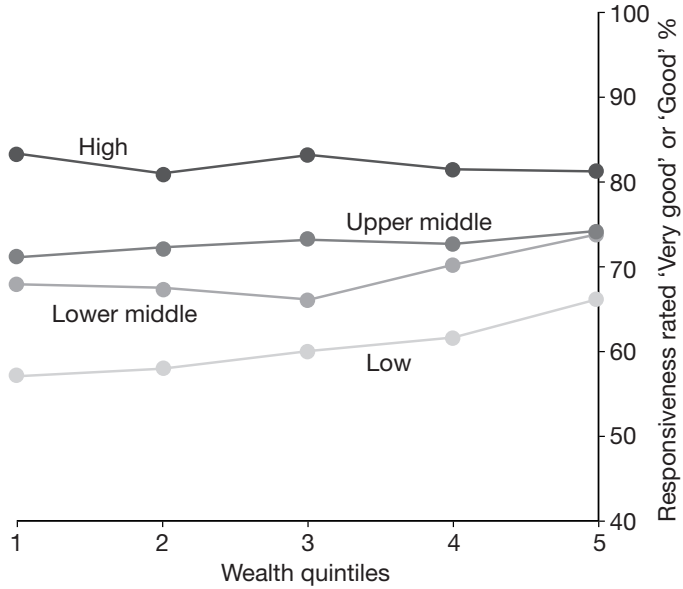

Fig. 2.5.13 Gradient in responsiveness for population groups within countries by wealth quintiles

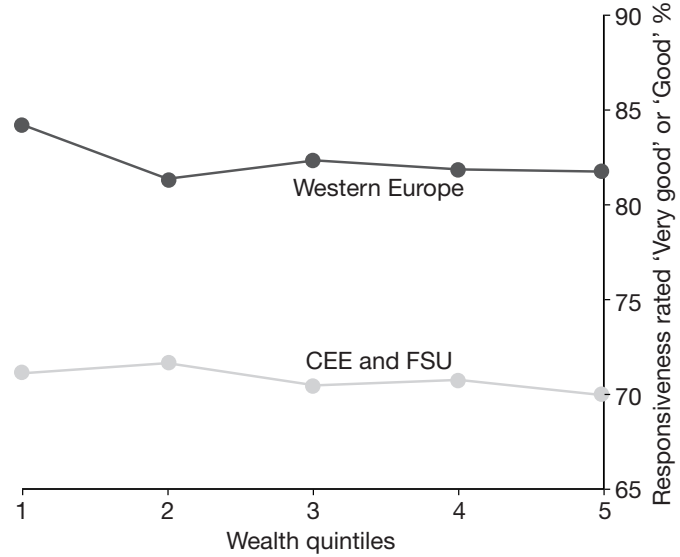

Fig. 2.5.14 Gradient in responsiveness for population groups within countries in Europe by wealth quintiles

completion of valid antenatal care as a means of understanding the relationship between responsiveness and coverage in general.

Keeping all other factors constant, well-resourced health system environments should be able to afford better quality care and receive better responsiveness ratings from users. Using a simple correlation for each responsiveness domain and keeping development contexts constant (by looking at correlations within World Bank country income groups), we observed whether higher health expenditures are associated with higher responsiveness and for which domains. Fig. 2.5.15 lists the domains for which the correlations between total and government health expenditures and responsiveness are significant (p=0.05). In general, there is a positive association across many of the domains for most country income groupings, with the exception of lower middle-income countries. This indicates that increases in health expenditures in this grouping of countries are not being translated into improvements in patients' experiences of care, perhaps because absolute levels of expenditure are too low to create even a basic health system.

Where particular health needs require multiple contacts with the health system (e.g. chronic conditions or treatment protocols for TB or maternal care), the interaction between provider and user behaviours can influence utilization patterns. Under- or incorrect utilization can influence technical care and health outcomes (Donabedian 1973).[8]

A few simple analyses of responsiveness and adherence-related data give a sense of the extent of validity in the WHS responsiveness results and how the acceptability and accessibility of services, as measured by responsiveness, can lead to adherence. Fig. 2.5.16 shows a scatterplot of responsiveness and antenatal coverage rates. The latter rates were obtained from the WHS question which asked whether the respondent had completed four antenatal visits. Overall, a significant linear correlation was observed between the level of responsiveness and the percentage of respondents reporting that they had completed all four antenatal visits (r=0.51, p=0.000). The highest correlations were observed for the level of dignity (r=0.55), communication (0.54) and confidentiality (0.50). The responsiveness measure of inequality was less strongly correlated (r=0.35).

[8] This assumes that, when applied technically correctly, health interventions have a positive impact on health.

Fig. 2.5.15 Correlations of average total health expenditure per capita and overall responsiveness for countries in different World Bank income categories

	AMBULATORY		INPATIENT	
	Total health expenditure per capita	**Government health expenditure per capita**	**Total health expenditure per capita**	**Government health expenditure per capita**
Low income (n, 19)	• Higher levels for basic amenities, confidentiality • Lower inequalities for dignity and autonomy	• Higher levels for basic amenities, dignity, confidentiality • Lower inequalities for dignity and basic amenities	• Higher levels for basic amenities • Lower inequalities for all domains except prompt attention.	• Higher levels for basic amenities • Lower inequalities for dignity
Lower-middle income (n, 15)	• None	• None	• None.	• Higher levels for dignity
Higher- middle income (n,12)	• Higher levels for communication, choice	• Higher levels for dignity, communication, choice	• Higher levels for choice, social support	• Higher levels for prompt attention, choice, social support
High income (n,15)	• Higher levels for communication, autonomy, choice, basic amenities. • Lower inequalities for basic amenities	• Higher levels for all domains except confidentiality. • Lower inequalities for basic amenities.	• Higher levels for all domains except communication and confidentiality.	• Higher levels for all domains except confidentiality • Lower inequalities for prompt attention, dignity, social support

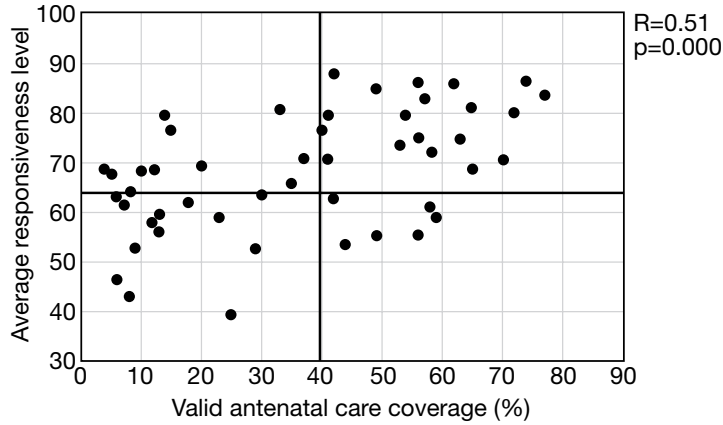

Fig. 2.5.16 Responsiveness and antenatal coverage

Conclusions

Empowering patients and equity in access are founding values that underpin the outlook for the new European health strategy. These values are expressed in the *White Paper: Together for Health: A Strategic Approach for the EU 2008-2013* (Commission of the European Communities 2007). Ensuring high responsiveness performance from health systems, with respect to both level and equity, is one key strategy to support these values. Measuring responsiveness is one approach to keeping the issue high on the health systems performance agenda.

The analyses for this chapter used inequalities in responsiveness across income groups as a proxy for inequities in responsiveness. The discussion below refers to these two aspects of responsiveness.

Common concerns

A wide array of results on health system responsiveness has been presented in this chapter. Health systems across the world show some common strengths and failings. Nurses' and doctors' respectful treatment of users is encapsulated in the responsiveness domain – dignity. This is a relative strength in comparison to systemic issues such as prompt attention, involvement in decision-making (autonomy) or choice (/continuity of provider).

Our analysis has generally confirmed the hypothesis of a positive relationship between a country's level of development (represented by national income) and the responsiveness of its health system (as is observed for health outcomes). However, while there is a linear relationship between the income level in a country and the average level of responsiveness, dramatic reductions in responsiveness inequalities are only observed in the high-income country category. This observation was true for both inpatient and ambulatory care.

Elevated levels of health expenditures are no guarantee that a system's responsiveness has improved. For lower middle-income countries no gains in responsiveness are observed for increases in health expenditures, probably due to inadequate general funding. Increased health expenditure (particularly in the public sector) for the other country groupings does yield gains in the overall responsiveness level and equality, but usually in some specific domains. On the other hand, lower responsiveness is associated with lower coverage and inequalities in responsiveness are associated with greater inequity in access, regardless of development setting. Hence, explicit steps are needed to build good levels of responsiveness performance into all systems.

The European analysis showed substantial differences in mean levels and within-country inequalities between western European and CEE/FSU countries. Average responsiveness levels are higher in western European (85%) than in CEE/FSU (73%) countries. In both groups of countries, ambulatory services had the highest levels for dignity and the highest inequalities for prompt attention. In inpatient services, levels of dignity were highest in both country groupings but prompt attention inequities were highest in CEE/FSU countries and autonomy and confidentiality inequalities were highest in western Europe.

Implementing change

Enhancing communication in the health system provides a potential entry point for improving responsiveness. Clear communication is associated with dignity, better involvement in decision-making and, in addition, supports better coverage or access. It is also an attribute that is highly valued by most societies. In the European context, it is interesting to note that CEE/FSU countries place special importance on communication (Valentine et al. 2008).

As shown here, responsiveness appears to be complementary or contributory to ensuring equity in access (to the technical quality of care). This is in keeping with the Aday and Andersen (1974) framework and with Donabedian (1980) who introduced the concept of the quality of health care and satisfaction with the care received as a valid component for achieving high technical quality of care and high rates of access to care. Inequities in access will result if the process of care systematically dissuades some groups from either initiating or continuing use of services to obtain the maximum benefit from the intervention. It is critical to deliver health interventions effectively and ensure compliance in primary care where a large majority of the population receives preventive and promotive health interventions. This is likely to become an increasing concern with the global epidemiological transition from infectious to chronic diseases. Therefore, primary-care providers need to be aware of their critical role in patient communication and treating individuals with respect.

Responsiveness measurement and future research

The psychometric properties of the responsiveness questions show resilience across different countries and settings and indicate that the responsiveness surveys (when reported as raw data) have face validity. The WHS managed to improve on the MCS Study questions in several ways and provides a useful starting tool for countries embarking on routine assessments of responsiveness.

Some key aspects of responsiveness still need to be researched further. In particular, while theoretically complementary, further investigation could benefit empirical research on the potential trade-offs between health (through investments in improved technical applications) and non-health (through better responsiveness) outcomes.

A second key area relates to gaining a better understanding of how responsiveness and responsiveness inequities may act as indicators of inequities in access or unmet need in the population and what measures can be taken to improve responsiveness in the light of this relationship.

A third key area relates to the self-reported nature of the responsiveness instrument. Self-reported data may be prone to measurement error (e.g. Groot 2000; Murray et al. 2001) where bias results from groups of respondents (for example defined by socio-economic charac-

teristics) varying systematically in their reporting of a fixed level of the measurement construct. The degree of comparability of self-reported survey data across individuals, socio-economic groups or populations has been debated extensively, usually with regard to health status measures (e.g. Bago d'Uva et al. 2007; Lindeboom & van Doorslaer 2004).

Similar concerns apply to self-reported data on health systems responsiveness where the characteristics of the systems and cultural norms regarding the use and experiences of public services are likely to predominate. The method of anchoring vignettes has been promoted as a means for controlling for systematic differences in preferences and norms when responding to survey questions (see Salomon et al. 2004). Vignettes represent hypothetical descriptions of fixed levels of a construct (such as responsiveness) and individuals are asked to evaluate these in the same way that they are asked to evaluate their own experiences of the health system. The vignettes provide a source of external variation from which information on systematic reporting behaviour can be obtained. To date, little use has been made of the vignette data within the WHS (Rice et al. 2008) and these offer a valuable area for future research.

Prospects for measuring responsiveness

Non-health outcomes are gaining increasing attention as valid measures of performance and quality. These require some feedback on what happens when users make contact with health-care systems and that can be easily compared across countries. Routine surveys on responsiveness are by no means a substitute for other forms of participation but, within the theme of patient empowerment, can provide opportunities for users' voices to be heard in health-care systems.

Responsiveness measurement (as opposed to broader patient satisfaction measurement) is increasingly recognized as an appropriate approach for informing health system policy. Work by the Picker Institute (1999) and the AHRQ (1999); the future work envisaged by the OECD (Garratt et al. 2008); and the broader analytical literature have built this case very satisfactorily. The work of the last decade has provided a solid base and an opportunity for individual countries to introduce measures of responsiveness into their health-policy information systems in the short and medium term.

References

Aday, LA. Andersen, R (1974). 'A framework for the study of access to medical care'. *Health Services Research*, 9(3): 208–220.

AHRQ (1999). *CAHPS 2.0 survey and reporting kit*. Rockville, MD: Agency for Healthcare Research and Quality.

Andersen, RM (1995). 'Revisiting the behavioral model and access to medical care: does it matter?' *Journal of Health and Social Behavior*, 36(1): 1–10.

Bago d'Uva, T. van Doorlsaer, E. Lindeboom, M. O Donnell, O (2007). 'Does reporting heterogeneity bias the measurement of health disparities?' *Health Economics*, 17(3): 351–375.

Blendon, RG. Schoen, C. DesRoches. C. Osborn, R. Zapert, K (2003). 'Common concerns amid diverse systems: health care experiences in five countries. The experiences and views of sicker patients are bellwethers for how well health care systems are working.' *Health Affairs,* 22(3): 106–121.

Bradley, EH. McGraw, SA. Curry, L. Buckser, A. King, KL. Kasl, SV. Andersen, R (2002). 'Expanding the Andersen model: the role of psychosocial factors in long-term care use.' *Health Services Research*, 37(5): 1221–1242.

Commission of the European Communities (2007). *White Paper. Together for health: a strategic approach for the EU 2008–2013*. Brussels: European Commission (http://ec.europa.eu/health/ph_overview/Documents/strategy_wp_en.pdf).

De Silva, A (2000). *A framework for measuring responsiveness*. GPE Discussion Paper Series No. 32 (http://www.who.int/responsiveness/papers/en).

Donabedian, A (1973). *Aspects of medical care administration*. Cambridge, MA: Harvard University Press.

Donabedian, A (1980). *Explorations in quality assessment and monitoring: the definition of quality and approaches to assessment*. Ann Arbor, Michigan: Health Administration Press.

Garratt, AM. Solheim, E. Danielsen, K (2008). *National and cross-national surveys of patient experiences: a structured review*. Oslo: Norwegian Knowledge Centre for the Health Services (Report No. 7).

Gilson, L. Doherty, J. Loewenson, R. Francis, V (2007). *Challenging inequity through health systems*. Final Report Knowledge Network on Health Systems (http://www.who.int/social_determinants/knowledge_networks/final_reports/en/index.htm).

Groot, W (2000). 'Adaptation and scale of reference bias in self-assessments of quality of life.' *Journal of Health Economics*, 19: 403–420.

Hall, JA. Feldstein, M. Fretwell, MD. Rowe, JW. Epstein, AM (1990). 'Older patients' health status and satisfaction with medical care in an HMO population.' *Medical Care*, 28: 261–70.

Harper, S. Lynch, J (2006). Measuring health inequalities. In: Oakes, JM. Kaufman, JS (eds.). *Methods in social epidemiology*. San Francisco: John Wiley & Sons.

Labarere, J. Francois, P. Auquier, P. Robert, C. Fourny, M (2001). 'Development of a French inpatient satisfaction questionnaire.' *International Journal for Quality in Health Care*, 13: 99–108.

Landis, JR. Koch, GG (1977). 'The measurement of observer agreement for categorical data.' *Biometrics*, 33: 159–174.

Lindeboom, M. van Doorslaer E (2004). 'Cut-point shift and index shift in self-reported health.' *Journal of Health Economics,* 23(6): 1083–1099.

Murray, CJL. Frenk, J (2000). 'A framework for assessing the performance of health systems.' *Bulletin of the World Health Organization*, 78: 717–731.

Murray, CJL. Tandon, A. Salomon, J. Mathers, CD (2001). *Enhancing cross-population comparability of survey results*. Geneva: WHO/EIP (GPE Discussion Paper No. 35).

Nunnally, JC. Bernstein, IH (1994). *Psychometric theory, 3rd ed*. New York: McGraw-Hill.

Picker Institute (1999). *The Picker Institute Implementation Manual*. Boston, MA: Picker Institute.

Rice, N. Robone, S. Smith, PC (2008). *The measurement and comparison of health system responsiveness*. Presentation to Health Econometrics and Data Group (HEDG), January 2008, University of Norwich (HEDG Working Paper 08/05).

Salomon, J. Tandon, A. Murray, CJ (2004). 'Comparability of self-rated health: cross-sectional multi-country survey using anchoring vignettes.' *British Medical Journal,* 328(7434): 258.

Shengelia, B. Tandon, A. Adams, O. Murray, CJL (2005). 'Access, utilization, quality, and effective coverage: an integrated conceptual framework and measurement strategy.' *Social Science & Medicine*, 61: 97–109.

Solar, O. Irwin, A (2007). *A conceptual framework for action on the social determinants of health*. Draft discussion paper for the Commission on Social Determinants of Health. April 2007 (http://www.who.int/social_determinants/resources/csdh_framework_action_05_07.pdf).

Steine, S. Finset, A. Laerum, E (2001). 'A new, brief questionnaire (PEQ) developed in primary health care for measuring patients' experience of interaction, emotion and consultation outcome.' *Family Practice*, 18(4): 410–419.

Tanahashi, T (1978). 'Health service coverage and its evaluation.' *Bulletin of the World Health Organization*, 56(2): 295–303.

Üstün, TB. Chatterji, S. Mechbal, A. Murray, CJL. WHS Collaborating Groups (2003). The world health surveys. In: Murray, CJL. Evans, DB (eds.). *Health systems performance assessment: debates, methods and empiricism*. Geneva: World Health Organization.

Üstün, TB. Chatterji, S. Mechbal, A. Murray, CJL (2005). Quality assurance in surveys: standards, guidelines and procedures. In: *Household surveys in developing and transition countries: design, implementation and analysis*. New York: United Nations (http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf).

Üstün, TB. Chatterji, S. Villanueva, M. Bendib, L. Çelik. C. Sadana, R. Valentine, N. Ortiz, J. Tandon, A. Salomon, J. Cao, Y. Jun, XW. Özaltin, E. Mathers, C. Murray, CJL (2001). *WHO multi-country survey study on health and responsiveness 2000–2001*. Geneva: World Health Organization (GPE Discussion Paper 37) (http://www.who.int/healthinfo/survey/whspaper37.pdf).

Valentine, N. Bonsel, GJ. Murray. CJL (2007). 'Measuring quality of health care from the user's perspective in 41 countries: psychometric properties of WHO's questions on health systems responsiveness.' *Quality of Life Research*, 16(7): 1107–1125.

Valentine, N. Darby, C. Bonel, GJ (2008). 'Which aspects of non-clinical quality of care are most important? Results from WHO's general population surveys of health systems responsiveness in 41 countries.' *Social Science and Medicine,* 66(9): 1939–1950.

Valentine, NB. de Silva, A. Kawabata, K. Darby, C. Murray, CJL. Evans, DB. (2003). Health system responsiveness: concepts, domains and operationalization. In: Murray, CJL. Evans, DB (eds.). *Health systems performance assessment: debates, methods and empiricism*. Geneva: World Health Organization.

Valentine, NB. Lavallee, R. Liu, B. Bonsel, GJ. Murray, CJL (2003a). Classical psychometric assessment of the responsiveness instrument in the WHO multi-country survey study on health responsiveness 2000–2001. In: Murray, CJL. Evans, DB (eds.). *Health systems performance assessment: debates, methods and empiricism*. Geneva: World Health Organization.

Ware, JE. Hays, RD (1988). 'Methods for measuring patient satisfaction with specific medical encounters.' *Medical Care*, 26(4): 393–402.

WHO (2000). *The world health report 2000. Health systems: improving performance*. Geneva: World Health Organization.

WHO (2001). Report on WHO meeting of experts responsiveness (HFS/FAR/RES/00.1) Meeting on Responsiveness Concepts and Measurement.

Geneva, Switzerland: 13–14 September 2001 (http://www.who.int/health-systems-performance/technical_consultations/responsiveness_report.pdf).

WHO (2005). *The health systems responsiveness analytical guidelines for surveys in the multi-country survey study*. Geneva: World Health Organization (http://www.who.int/responsiveness/papers/MCSS_Analytical_Guidelines.pdf).

WHO (2005a). *WHO glossary on social justice and health*. A report of the WHO Health and Human Rights, Equity, Gender and Poverty Working Group. Available online at WHO, forthcoming.

WHO & EQUINET (forthcoming). *A framework for monitoring equity in access and health systems strengthening in AIDS treatment programmes: options and implementation issues*. Geneva: World Health Organization and EQUINET.

Annex 1

Groupings of World Health Survey countries

Fig. A WHS countries grouped by World Bank income categories

Low income Bangladesh, Burkina Faso, Chad, Comoros, Congo, Cote d'Ivoire, Ethiopia, Ghana, India, Kenya, Lao People's Democratic Republic, Malawi, Mali, Mauritania, Myanmar, Nepal, Pakistan, Senegal, Viet Nam, Zambia, Zimbabwe	**Lower-middle income** Bosnia and Herzegovina, Brazil, China, Dominican Republic, Ecuador, Georgia, Guatemala, Kazakhstan, Morocco, Namibia, Paraguay, Philippines, Sri Lanka, Tunisia, Ukraine
Higher-middle income Croatia, Czech Republic, Estonia, Hungary, Latvia, Malaysia, Mauritius, Mexico, Russian Federation, Slovakia, South Africa, Uruguay	**High income** Austria, Belgium, Denmark, Finland, Germany, Greece, Ireland, Israel, Italy, Luxembourg, Netherlands, Portugal, Slovenia, Spain, Sweden, United Arab Emirates, United Kingdom

Fig. B WHS countries in Europe

CEE/FSU Bosnia and Herzegovina, Croatia, Czech Republic, Estonia, Georgia, Hungary, Kazakhstan, Latvia, Russia, Slovakia, Slovenia, Ukraine	**Western Europe** Austria, Belgium, Denmark, Finland, Germany, Greece, Ireland, Israel, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom

Annex 2 *WHS 2002 sample descriptive statistics*

Country	Response rate - interview completion (%)	Users of ambulatory services in last twelve months	Users of inpatient services in last three years	Percentage female	Average age (years)	Percentage high school or more educated	Percentage in good or very good health
Low income							
Bangladesh	85	4020	777	53	39	8	44
Burkina Faso	96	1199	589	53	36	3	70
Chad	92	423	371	53	37	3	58
Comoros	95	526	374	55	42	5	54
Congo	79	381	288	53	36	18	56
Cote d'Ivoire	97	765	305	43	36	13	60
Ethiopia	96	1779	224	52	37	3	75
Ghana	70	1567	677	55	41	4	72
India	93	5003	1735	51	39	21	58

Annex 2 *cont'd*

Country	Response rate - interview completion (%)	Users of ambulatory services in last twelve months	Users of inpatient services in last three years	Percentage female	Average age (years)	Percentage high school or more educated	Percentage in good or very good health
Kenya	82	2228	803	58	38	21	66
Lao People's Democratic Republic	98	735	570	53	38	10	78
Malawi	93	2423	1236	58	36	1	79
Mali	79	130	104	43	42	3	70
Mauritania	98	552	469	61	39	10	69
Myanmar	97	1667	320	57	41	9	79
Nepal	98	3279	1141	57	39	5	62
Pakistan	93	3727	913	44	37	14	75
Senegal	88	222	182	48	38	8	58
Viet Nam	84	1541	548	54	40	24	51
Zambia	88	2188	764	55	36	5	72

Zimbabwe	94	1660	649	64	37	5	52
Lower-middle income							
Bosnia and Herzegovina	94	394	259	58	47	8	58
Brazil	100	2341	1244	56	42	28	53
China	100	1435	423	51	45	28	62
Dominican Republic	74	1315	1508	54	42	5	56
Ecuador	77	1372	592	56	41	13	57
Georgia	92	763	227	58	49	88	38
Guatemala	98	2063	978	62	40	12	53
Kazakhstan	100	2331	803	66	41	96	48
Morocco	79	2211	800	59	41	14	41
Namibia	91	650	862	59	38	4	72
Paraguay	97	2414	1096	54	40	12	70

Annex 2 *cont'd*

Country	Response rate - interview completion (%)	Users of ambulatory services in last twelve months	Users of inpatient services in last three years	Percentage female	Average age (years)	Percentage high school or more educated	Percentage in good or very good health
Philippines	100	2625	906	52	39	16	60
Sri Lanka	99	2268	1697	53	41	21	72
Tunisia	96	2352	816	53	42	28	62
Ukraine	99	735	580	64	48	87	27
Upper-middle income							
Croatia	100	465	259	59	52	16	51
Czech Republic	49	411	302	55	48	47	55
Estonia	99	395	289	64	50	74	36
Hungary	100	453	489	58	49	63	51
Latvia	92	283	293	67	51	34	33

Malaysia	80	1943	1329	56	41	42	78
Mauritius	88	1702	1180	52	42	13	65
Mexico	97	19457	1440	55	42	23	67
Russian Federation	100	1794	1019	64	51	61	31
Slovakia	99	897	355	62	39	71	66
South Africa	89	384	384	53	38	34	73
Uruguay	100	1029	536	51	46	30	79
High income							
Austria	100	184	351	62	45	26	77
Belgium	100	298	299	56	45	64	74
Denmark	100	316	194	53	51	52	79
Finland	100	464	345	55	53	58	55
Germany	100	428	401	60	50	23	65
Greece	100	433	272	50	51	47	67

Annex 2 *cont'd*

Country	Response rate - interview completion (%)	Users of ambulatory services in last twelve months	Users of inpatient services in last three years	Percentage female	Average age (years)	Percentage high school or more educated	Percentage in good or very good health
Ireland	100	239	214	55	44	19	82
Israel	57	521	412	57	45	85	76
Italy	100	541	232	57	48	51	63
Luxembourg	100	135	237	52	45	43	73
Netherlands	100	624	192	67	44	83	76
Portugal	100	510	212	62	50	20	39
Slovenia	44	284	72	53	47	52	58
Spain	53	2863	1601	41	53	31	64
Sweden	100	300	266	58	51	70	62
United Arab Emirates	100	453	239	48	37	65	86
United Kingdom	100	369	344	63	50	46	68

2.6 *Measuring equity of access to health care*

SARA ALLIN, CRISTINA HERNÁNDEZ-QUEVEDO, CRISTINA MASSERIA

Introduction

A health system should be evaluated against the fundamental goal of ensuring that individuals in need of health care receive effective treatment. One way to evaluate progress towards this goal is to measure the extent to which access to health care is based on need rather than willingness or ability to pay. This egalitarian principle of equity or fairness is the primary motivation for health systems' efforts to separate the financing from the receipt of health care as expressed in many policy documents and declarations (Judge et al. 2006; van Doorslaer et al. 1993). The extent to which equity is achieved is thus an important indicator of health system performance.

Measuring equity of access to care is a core component of health system performance exercises. The health system performance framework developed in WHO's *The world health report 2000* stated that ensuring access to care based on need and not ability to pay is instrumental in improving health (WHO 2000). It can also be argued that access to care is a goal in and of itself: 'beyond its tangible benefits, health care touches on countless important and in some ways mysterious aspects of personal life and invests it with significant value as a thing in itself' (President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioural Research, 1983 cited in Gulliford et al. 2002). Equitable access to health care has been identified as a key indicator of performance by the OECD (Hurst & Jee-Hughes 2001) and underlies European-level strategies such as those developed at the European Union Lisbon summit in March 2000 and the Open Method of Coordination for social protection and social inclusion (Atkinson et al. 2002).

However, it is far from straightforward to measure equity and translate such measures into policy. This chapter is structured according to three objectives: (i) to review the conceptualization and measurement of equity in the health system, with a focus on access to care; (ii) to present the strengths and weaknesses of the common methodological approaches to measuring equity, drawing on illustrations from the existing literature; and (iii) to discuss the policy implications of equity analyses and outline priorities for future research.

Defining equity, access and need

Libertarianism and egalitarianism are two ideological perspectives that dominate current debates about individuals' rights to health care (Donabedian 1971; Williams 1993; Williams 2005). Libertarians are concerned with preserving personal liberty and ensuring that minimum health-care standards are achieved. Moreover, access to health care can be seen as a privilege and not a right: people who can afford to should be able to pay for better or more health care than their fellow citizens (Williams 1993). Egalitarians seek to ensure that health care is financed according to ability to pay and delivery is organized so that everyone has the same access to care. Care is allocated on the basis of need rather than ability to pay, with a view to promote equality in health (Wagstaff & van Doorslaer 2000). Egalitarians view access to health care as a fundamental human right that can be seen as a prerequisite for personal achievement, therefore it should not be influenced by income or wealth (Williams 1993).

These debates are also informed by the comprehensive theory of justice developed by Rawls (1971) that outlines a set of rules which would be accepted by impartial individuals in the 'original position'. This original position places individuals behind a 'veil of ignorance' – having no knowledge of either their place in society (social standing) or their level of natural assets and abilities. The Rawlsian perspective has been interpreted to suggest that equity is satisfied if the most disadvantaged in society have a decent minimum level of health care (Williams 1993). This would be supported by libertarians provided that government involvement was kept to a minimum. However, if

health care is considered one of Rawls' social primary goods[1] then an equitable society depends on the equal distribution of health care, in line with egalitarian goals. Furthermore, to the extent that health care can be considered essential for individuals' capability to function, then the egalitarian perspective is also consistent with Sen's theory of equality of capabilities (Sen 1992).

No perfectly libertarian or egalitarian health system exists but the egalitarian viewpoints are largely supported by both the policy community and the public. This support is evidenced by the predominantly publicly funded health systems with strong government oversight that separate payment of health care from its receipt and offer programmes to support the most vulnerable groups. At international level the view that access to health care is a right is illustrated by the 2000 Charter of Fundamental Rights of the European Union and the 1948 Universal Declaration of Human Rights.

The debate between libertarian and egalitarian perspectives is not resolved in practice. Policies that preserve individual autonomy and freedom of choice exist alongside policies of redistribution, as evidenced by the existence of a private sector in health care that allows those able or willing to pay to purchase additional health services. Thus the design of the health system impacts equity of access to health care. For instance, patient cost sharing may introduce financial barriers to access for poorer populations and voluntary health insurance may allow faster access or access to better quality services for the privately insured (Mossialos & Thomson 2003). Policy-makers appear to be concerned about the effects of health-care financing arrangements on the distribution of income and the receipt of health care (OECD 1992; van Doorslaer et al. 1993). Chapter 2.4 on financial protection provides an in-depth review of the extent to which health systems ensure that the population is protected from the financial consequences of accessing care.

[1] Social primary goods are those that are important to people but created, shaped and affected by social structures and political institutions. These contrast with the natural primary goods (intelligence, strength, imagination, talent, good health) that inevitably are distributed unequally in society (Rawls 1971).

What objective of equity do we want to evaluate?

The idea that health systems should pursue equity goals is widely supported. However, it is not straightforward to operationalize equity in the context of health care. Many definitions of equity in health-care delivery have been debated and Mooney identifies seven in the economics literature (Mooney 1983 & 1986). The first two (equality of expenditure per capita, equality of inputs across regions) are unlikely to be equitable since they do not allow for variations in levels of need for care. The third (equality of input for equal need) accounts for need but does not consider factors that may give rise to inequity beyond the size of the health-care budget. The fourth and fifth are the most commonly cited definitions – equality of access for equal need (individuals should face equal costs of accessing care) and equality of utilization for equal need (individuals in equal need should not only face equal costs but also demand the same amount of services). The sixth suggests that if needs are prioritized/ranked in the same way across regions, then equity is achieved when each region is just able to meet the same 'last' or 'marginal' need. The seventh argues that equity is achieved if the level of *health* is equal across regions and social groups, requiring positive discrimination in favour of poorer people/regions and an unequal distribution of resources.

All the above goals are concerned with health-care delivery. Equity in health care is often defined in terms of health-care financing whereby individuals' payments for health care should be based on their ability to pay and therefore proportional to their income. Individuals with higher incomes should pay more and those with lower incomes should pay less, regardless of their risk of illness or receipt of care. This concept is based on the vertical equity principle of unequal payment for unequals in which unequals are defined in terms of their level of income (Wagstaff & van Doorslaer 2000; Wagstaff et al. 1999). It has direct implications for access to care since financial barriers to access may arise from inequitable (or regressive) systems of health-care finance. The financial arrangements of the health system not only impact on equity of access to health care but also have the potential to exacerbate health inequalities: "unfair financing both enhances any existing unfairness in the distribution of health and compounds it by making the poor multiply deprived" (Culyer 2007, p.15).

The policy perspective requires a working definition of equity that is feasible (i.e. within the scope of health policy) and makes intuitive sense. In an attempt to clarify equity principles for policy-makers, Whitehead (1991) builds on Mooney's proposed equity principles to develop an operational definition encompassing the three dimensions of accessibility, acceptability and quality.

1. Equal access to available care for equal need – implies equal entitlements (i.e. universal coverage); fair distribution of resources throughout the country (i.e. allocations on basis of need); and removal of geographical and other barriers to access.
2. Equal utilization for equal need – to ensure use of services is not restricted by social or economic disadvantage (and ensure appropriate use of essential services). This accepts differences in utilization that arise from individuals exercising their right to use or not use services according to their preferences. This is consistent with the definition of equity that is linked to personal choice, such that an outcome is equitable if it arises in a state in which all people have equal choice sets (Le Grand 1991).
3. Equal quality of care for all – implies an absence of preferential treatments that are not based on need; same professional standards for everyone (for example, consultation time, referral patterns); and care that is considered to be acceptable by everyone.

In a similar exercise to identify an operational definition of equity that is relevant to policy-makers and aligned with policy objectives, equal access for equal need is argued to be the most appropriate definition because it is specific to health care and respects the potentially acceptable reasons for differentials in health-care utilization (Oliver & Mossialos 2004). Moreover, unequal access across groups defined by income or socio-economic status is the most appropriate starting point for directing policy and consistent with many governments' aims to provide services on the basis of need rather than ability to pay (Oliver & Mossialos 2004).

The goal of equal (or less unequal) health outcomes appears to be shared by most governments, as expressed in policy statements and international declarations (such as European Union's Health and Consumer Protection Strategy and Programme 2007-2013; WHO's Health 21 targets) (Judge et al. 2006). However, two factors complicate the adoption of equality in health to evaluate health-care performance.

First, social and economic determinants of health fall outside the health system and beyond the scope of health policy and health care. Second, such an action might require restrictions on the ways in which people choose to live their lives (Mooney 1983). In the 1990s the policy support for improving equity of access or receipt of care was more evident than the commitment to improve equality in health (Gulliford 2002). However, more recently the reduction of avoidable health inequalities has become a priority government objective in the United Kingdom (Department of Health 2002 & 2003). The formula used to allocate resources to the regions seeks to improve equity in access to services and to reduce health inequalities (Bevan 2008).

These two principles are clearly linked. Much support for the equity objective based on access derives from its potential for achieving equality in health. Some argue that an equitable distribution of health leads to a more equal distribution of health (Culyer & Wagstaff 1993). Health care is instrumental in improving health or minimizing ill-health. In fact, no one wants to consume health care in a normal situation but it becomes essential at the moment of illness. Demand for health care is thus derived from the demand for health itself (Grossman 1972). Ensuring an equitable distribution of health-care resources serves a broader aim of health improvement and reduction of health inequalities. From the egalitarian viewpoint it is often argued that allocating health-care resources according to need will promote, if not directly result in, equality in health (Wagstaff & van Doorslaer, 2000). Culyer and Wagstaff (1993) demonstrate that this is not necessarily the case but Hurley argues that equality of access is based on the ethical notion of equal opportunity or a fair chance and not necessarily on the consequences of such access, such as utilization or health outcomes (Hurley 2000).

How to define access?

The equity objective of equal access for equal need commands general policy support but the questions of how to define and measure access need to be clarified. Narrowly defined, access is the money and time costs people incur obtaining care (Le Grand 1982; Mooney 1983). One definition of access incorporates additional dimensions: 'the ability to secure a specified set of health care services, at a specified level of quality, subject to a specified maximum level of personal

inconvenience and cost, whilst in possession of a specified amount of information' (Goddard & Smith 2001, p.1151).

Accessing health care depends on an array of supply- and demand-side factors (Healy & McKee, 2004). Supply-side factors that affect access to and receipt of care include the volume and distribution of human resources and capital; waiting times; referral patterns; booking systems; how individuals are treated within the system (continuity of care); and quality of care (Gulliford et al. 2002b; Starfield, 1993; Whitehead, 1991). The demand-side has predisposing, enabling and needs factors (Aday & Andersen, 1974), including socio-demographics; past experiences with health care; perceived quality of care; perceived barriers; health literacy; beliefs and expectations regarding health and illness; income levels (ability to pay); scope and depth of insurance coverage; and educational attainment.

The complexity of the concept of access is apparent in the multitude of factors that affect access and potential indicators of access. As a result, many researchers use access synonymous with utilization, implying that an individual's use of health services is proof that he/she can access these services. However, the two are not equivalent (Le Grand 1982; Mooney 1983). As noted, access can be viewed as opportunities available but receipt of treatment depends on both the existence of these opportunities and whether an individual actually makes use of them (Wagstaff & van Doorslaer 2000). Aday and Andersen suggest that a distinction must be made between 'having access' and 'gaining access' – the *possibility* of using a service if required and the actual *use* of a service, respectively (Aday & Andersen 1974; Aday & Andersen 1981). Similarly, Donabedian (1972, p. 111) asserts that: 'proof of access is use of service, not simply the presence of a facility' and thus it is argued that utilization represents *realized* access. In order to evaluate whether an individual has gained access, this view requires measurement of the actual utilization of health care and possibly also the level of satisfaction with that contact and health improvement.

A consensus about the most appropriate metric of access remains to be found. Many different elements or indicators of access can be measured (e.g. waiting time, availability of resources, access costs) and utilization can be directly observed. Therefore, while 'equal access for equal need' is arguably the principle of equity most appropriate for policy, 'equal utilization for equal need' is what is commonly measured and analysed. In this way, inequity is assumed to arise when

individuals in higher socio-economic groups are more likely to use or are using a greater quantity of health services after controlling for their level of need (see section below on defining need). However, it should be remembered that differences in utilization levels by socio-economic status (adjusting for need) do not necessarily imply inequity because they may be driven in part by individuals' informed choices or preferences (Le Grand 1991; Oliver & Mossialos 2004). Also an apparently equal distribution of needs-adjusted utilization by socio-economic status may not imply equity if the services used are low quality or inappropriate (Thiede et al. 2007).

Equity of access to health care could also be assessed directly by measuring the extent to which individuals did not receive the health care needed. Unmet need could be measured with clinical information (e.g. medical records or clinical assessments) or by self-report. Subjective unmet need is easily measurable and has been included in numerous recent health surveys e.g. European Union Statistics on Income and Living Conditions (EU-SILC) and the Survey of Health, Ageing and Retirement in Europe (SHARE). Levels of subjective unmet need and the stated reasons for unmet need could provide some insight into the extent of inequity in the system, particularly if these measures are complemented by information on health-care utilization.

How to define need?

An operational definition of need is required in order to examine the extent to which access or utilization is based upon it. Four possible definitions have been proposed in the economics literature (Culyer & Wagstaff 1993).

1. Need is defined in terms of an individual's current health status.
2. Need is measured by capacity to benefit from health care.
3. Need represents the expenditure a person ought to have i.e. the amount of health care required to attain health.
4. Need is indicated by the minimum amount of resources required to exhaust capacity to benefit.

The authors argue that the first definition is too narrow since it may miss the value of preventive care and certain health conditions may not be treatable (Culyer & Wagstaff, 1993). The second does not take account of the amount of resources spent or establish how much

health care a person needs. The third takes this into consideration since need is defined as the amount of health care required to attain equality of health. The fourth definition implies that when capacity to benefit is (at the margin) zero then need is zero; when there is positive capacity to benefit need is assessed by considering the amount of expenditure required to reduce capacity to benefit to zero (Culyer & Wagstaff 1993). However, by combining the level of need with the level of required resources the latter definition implies than an individual requiring more expensive intervention has greater need than someone with a potentially more urgent need but for less expensive treatment (Hurley 2000).

The definition of need as the capacity to benefit commands the widest approval in the economics literature (Folland et al. 2004). However, empirical studies measure need by level (and risk) of ill-health partly because of data availability and relative ease of measurement. The assumption that current health status reflects needs is generally considered to be reasonable – an individual in poor general health with a chronic condition clearly needs more health care than an individual in good health with no chronic condition. Also, individuals with higher socio-economic status have been shown generally to have more favourable prospects for health and thus greater capacity to benefit (Evans 1994) therefore allocation according to capacity to benefit may distort the allocation of resources away from the most vulnerable population groups. These latter groups would have worse ill health and allocating resources according to this principle would exacerbate socio-economic inequalities in health (Culyer 1995). From a utilitarian perspective, and to maximize efficiency, resources should be distributed in favour of those with the greatest capacity to benefit. However, an egalitarian perspective would conflict with the capacity to benefit definition of need because of the potential unintended implications for health inequality.

To measure need for health care, an individual's level of ill health is most commonly captured by a subjective measure of self-assessed health (SAH). This provides an ordinal ranking of perceived health status and is often included in general socio-economic and health surveys at European (e.g. European Community Household Panel; EU-SILC) and national level (e.g. British Household Panel Survey). The usual health question asks the respondent to rate their general health and sometimes includes a time reference (rate your health in the last twelve

months) or an age benchmark (compare your current health to individuals of your own age). Five categories are usually available for the respondent, ranging from very good or excellent to poor or very poor. SAH has been used extensively in the literature and has been applied to measure the relationship between health and socio-economic status (Adams et al. 2003); the relationship between health and lifestyles (Kenkel 1995); and the measurement of socio-economic inequalities in health (van Doorslaer et al. 1997).

Numerous methodological problems are associated with relying on SAH as a measure of need. An obvious concern relates to its reliability as a predictor of objective health status, but this may be misplaced. An early study from Canada found SAH to be a stronger predictor of seven-year survival among older people than their medical records or self-reports of medical conditions (Mossey & Shapiro 1982). This finding has been replicated in many subsequent studies and countries, showing that this predictive power does not vary across jurisdictions or socio-economic groups (Idler & Benyamini 1997; Idler & Kasl 1995). In their review of the literature, Idler and Benyamini (1997) argue that self-rated health represents an invaluable source of health status information and suggest several possible interpretations for its strong predictive effect on mortality.

- SAH measures health more accurately because it captures all illnesses a person has and possibly as yet undiagnosed symptoms; reflects judgements of severity of illness; and/or reflects individuals' estimates of longevity based on family history.
- SAH not only assesses current health but is also a dynamic evaluation thus representing a decline or improvement in health. Poor assessments of health may lessen an individual's engagement with preventive or self care or provoke non-adherence to screening recommendations, medications or treatments.
- SAH reflects social or individual resources that can affect health or an individual's ability to cope with illness.

Since this review, mounting evidence shows SAH to be a valid summary measure of health. It relates to other health-related indicators and appears to capture the broader influences of mortality (Bailis et al. 2003; Mackenbach et al. 2002; McGee et al. 1999; Singh-Manoux et al. 2006; Sundquist & Johansson, 1997); health-care use (van Doorslaer et al. 2000); and inequalities in mortality (van Doorslaer & Gerdtham 2003).

Self-assessed measures can be further differentiated into subjective and quasi-objective indicators (Jürges 2007), the latter based on respondents' reporting on more factual items such as specific conditions or symptoms. These quasi-objective indicators include the presence of chronic conditions (where specific chronic conditions are listed); specific types of cancer; limitations in activity of daily living (ADL) such as walking, climbing the stairs, etc; or in instrumental activity of daily living (IADL) such as eating or having a bath.

There is strong evidence that SAH is not only predictive of mortality and other objective measures of health but may be a more comprehensive measure of health status than other measures. However, bias is possible if different population groups systematically under- or over-report their health status relative to other groups. The subjective nature of SAH means that it can be influenced by a variety of factors that impact perceptions of health. Bias may arise if the mapping of true health in SAH categories varies according to respondent characteristics. Indeed, subgroups of the population appear to use systematically different cut-point levels when reporting SAH, despite equal levels of true health (Hernández-Quevedo et al. 2008). Moreover, the rating of health status is influenced by culture and language (Angel & Thoits 1987; Zimmer et al. 2000); social context (Sen 2002); gender and age (Groot 2000; Lindeboom & van Doorslaer 2004); and fears and beliefs about disease (Barsky et al. 1992). It is also affected by the way a question is asked e.g. the ordering of the question with other health-related questions or form-based rather than face-to-face interviews (Crossley & Kennedy 2002). Potential biases of SAH include state-dependence reporting bias (Kerkhofs & Lindeboom 1995); scale of reference bias (Groot 2000); and response category cut-point shift (Sadana et al. 2000).

Various approaches have been developed to correct for reporting bias in the literature. The first is to condition on a set of objective indicators of health and assume that any remaining variation in SAH reflects reporting bias. For example, Lindeboom and van Doorslaer (2004) use Canadian data and the McMaster Health Utilities Index as their quasi-objective measure of health. They find some evidence of reporting bias by age and gender but not for income. However, this approach relies on having a sufficiently comprehensive set of objective indicators to capture the variation in true health. The second approach uses health vignettes such as those in the current WHS (Bago d'Uva et

al. 2008). The third approach examines biological markers of disease risk in the countries considered for comparison, for example by combining self-reported data with biological data (Banks et al. 2006). Bias in reporting may affect estimates of inequalities. For example Johnston et al. (2007) report that the income gradient appears significant when using an objective measure of hypertension measured by a nurse as opposed to the self-reported measure of hypertension included in the Health Survey for England (HSE).

The availability of objective measures of health, such as biomarkers, is mostly limited to specific national surveys. At the European level, both the ECHP and EU-SILC include only self-reported measures. Only SHARE and the forthcoming European Health Interview Survey include some objective (e.g. walking speed, grip strength) and quasi-objective (e.g. ADL, symptoms) measures of health. At national level, only a few countries include objective measures, such as Finland (blood tests and anthropometric tests – FINRISK), Germany (anthropometric measures – National Health Interview and Examination Survey; urine and blood samples – German Health Survey for Children and Adolescents) and the United Kingdom – English Longitudinal Study of Ageing (ELSA) and HSE.

Biomarkers thus have limited availability and may still be subject to bias. The main methodological challenge lies with the standardization of data collection, as variations may arise from different methods. For example, a person's blood pressure may vary with the time of day. Often detailed information on data collection methods is not provided. This type of measurement error is particularly problematic if it is correlated with socio-demographic characteristics and hence biases estimates of social inequalities. Moreover, the collection of biological data also tends to reduce survey response rates, limiting sample size and representativeness (Masseria et al. 2007).

Overall, there is widespread support for equity goals in health care. However, no single operational definition of equity can capture the multiple supply- and demand-side factors that affect the allocation of effective, high-quality health care on the basis of need. This complexity necessitates not only a comprehensive set of information on individuals, their contacts with health care and system characteristics, but also on strong methodological techniques to assess these relationships empirically.

Methods for equity analysis

Methods of measuring equity of access to health care originated with comparisons of health-care use and health-care need (Collins & Klein 1980; Le Grand 1978) and have since taken broadly two directions. The first uses regression models to measure the independent effect of some measure of socio-economic status on the likelihood of contact with health services, the volume of health services used or the expenditures incurred (regression method). The second quantifies inequity by comparing the cumulative distribution of utilization with that of needs-adjusted utilization (ECuity method). Alternative metrics of equity are listed in Table 2.6.1.

Regression method

Regression analyses are the most commonly used means of measuring equity in the literature. These studies often draw on the behavioural model of health service use that suggests that health-care service use is a function of an individual's predisposition to use services (social structure, health beliefs); factors which enable or impede use on an individual (income and education) and community level (availability of services); and the level of need for care (Andersen 1995). Inequity thus arises when factors other than needs significantly affect the receipt of health care.

Regression models of utilization address the question – When needs and demographic factors affecting utilization are held constant, are individuals with socio-economic advantage (e.g. through income, education, employment status, availability of private insurance, etc.) more likely to access health care, and are they making more contacts, than individuals with less socio-economic advantage? A comprehensive model of utilization with multiple explanatory variables allows policy-relevant interpretations that can identify the factors that affect utilization and, to the extent that they are mutable, develop policies accordingly.

In the empirical literature, the most comprehensive studies of health service utilization have included explanatory variables that consider factors that capture not only needs but also individual predisposition and ability to use health-care services. Several studies of equity

Table 2.6.1 ***Examples of summary measures of socio-economic inequalities in access to health care***

Index	**Interpretation**
Correlation and regression	
Product-moment correlation	Correlation between health care utilization rate and socio-economic status (SES)
Regression on SES	Increase in utilization rate per one unit increase in SES
Regression on cumulative percentiles (relative index of inequality; Slope index of inequality)	Utilization rate ratio (RI/I) or differences (SII) between the least and most advantaged person
Regression on z-values	Utilization rate difference between group with lower and higher than average morbidity rates (x 0.5)
Gini-type coefficients	
Pseudo-Gini coefficient	0 = no utilization differences between groups; l = all utilization in hands of one person
Concentration index	0 = no utilization differences associated with SES; -1/+1 = all utilization in hands of least/most advantaged person
Horizontal inequity index	0 = no utilization differences associated with SES after need standardization; -1/+1 = all need standardized utilization in hands of least/most advantaged person
Generalized concentration index	Based on CI, but includes also mean distribution of health care

Source: adapted from Mackenbach & Kunst 1997

based on regression models have been conducted (Abásolo et al. 2001; Buchmueller et al. 2005; Dunlop et al. 2000; Häkkinen & Luoma 2002; Morris et al. 2005; Van der Heyden et al. 2003).

The study described here illustrates the methodology (Morris et al. 2005). The authors measured inequity in general practitioner consultations, outpatient visits, day cases and inpatient stays in England

between 1998 and 2000. A variety of need indicators were used, including not only age and gender but also self-reported indicators such as SAH; detailed self-reported indicators such as type of long-standing illness and GHQ-12 score; and ward-level health indicators including under-75 standardized mortality ratios and under-75 standardized illness ratios. Non-need variables such as income, education, employment status, social class and ethnicity were included. The effect of supply variables such as the Index of Multiple Deprivation access domain score, average number of general practitioners per 1000 inhabitants and average distance to acute providers were also considered, although their classification as needs or non-needs indicators is not straightforward (Gravelle et al. 2006; Morris et al. 2005).

The regression models showed that indicators of need were significantly associated with all health-care services (Table 2.6.2). People in worse health conditions were more likely to consult a general practitioner, to utilize outpatient and day care and to be hospitalized. However, non-need variables also played a significant role in determining access to health care (holding all else constant) which signalled inequity. Table 2.6.2 reports the marginal effects on utilization caused by income, education, ethnicity and supply. For example, people with higher incomes were significantly more likely to have an outpatient visit, those with lower educational attainment had a higher probability of consulting a general practitioner and education significantly affected the use of outpatient services. Distance and waiting time effects on utilization were also found.

This study provides an example of how regression models offer a rigorous and meaningful method of understanding the role of various socio-economic and system factors that affect access to health care within a country. However, this approach does not lend itself easily to cross-country and inter-temporal comparisons.

The ECuity method: concentration index

The ECuity method makes use of a regression model but tests for the existence of inequity by creating a relative index that allows comparisons across jurisdictions, time or sectors (O'Donnell et al. 2008). This method derives from the literature on income inequality based on the Lorenz curve and Gini index of inequality. While the Lorenz curve describes the distribution of income in a population, the

Table 2.6.2 ***Effect of specific non-need variables on health-care utilization, marginal effects***

	GP	Outpatient	Day cases	Inpatient
Ln (income)	-0.005	**0.011**	0.002	0.003
Education				
Higher education	0.007	0.023	0.001	**0.014**
A level or equivalent	0.014	0.009	-0.001	0.005
GCSE or equivalent	**0.014**	**0.020**	0.001	0.008
CSE or equivalent	**0.021**	0.021	0.008	0.004
Other qualifications	**0.032**	**0.041**	0.000	0.003
No qualifications	**0.015**	-0.003	-0.006	0.000
Ethnic group				
Black Caribbean	-0.006	-0.011	0.010	-0.009
Black African	0.009	-0.007	0.013	0.013
Black other	0.057	0.019	0.006	-0.016
Indian	**0.030**	-0.009	-0.009	-0.002
Pakistani	0.022	**-0.065**	-0.016	0.004
Bangladeshi	0.029	**-0.085**	0.015	-0.020
Chinese	-0.014	**-0.122**	-0.020	**-0.039**
Other non-white	0.012	**-0.043**	-0.002	0.014
Supply				
Access domain score	**-0.011**			
Proportion of outpatient <26 weeks		**0.351**		
GPs per 1000 patients			0.021	
Average distance to acute providers				**-0.0004**

Numbers in bold are statistically significant with 95% confidence interval

Source: Morris et al. 2005

concentration curve describes the relationship between the cumulative proportion of the population ranked by income (x-axis) and the cumulative proportion of health-care utilization (y-axis). Like the Gini index that provides a measure of income inequality, the concentration index is a measure of income-related inequality in access to health care and is estimated as twice the area between the concentration curve and the line of perfect equality (diagonal).

The concentration curves for actual medical care utilization (*LM*) and for needs-adjusted utilization (*LN*) are shown in Fig. 2.6.1. Individuals are ranked by a socio-economic variable (e.g. income) from the lowest or poorest to the highest or richest individual. If the cumulative proportion of both health-care utilization and needs-adjusted utilization are distributed equally across income then the two curves will coincide with the diagonal (line of perfect equality). If they lie above (below) the diagonal, the receipt of health care and the distribution of health-care need advantage the lower (higher) socio-economic

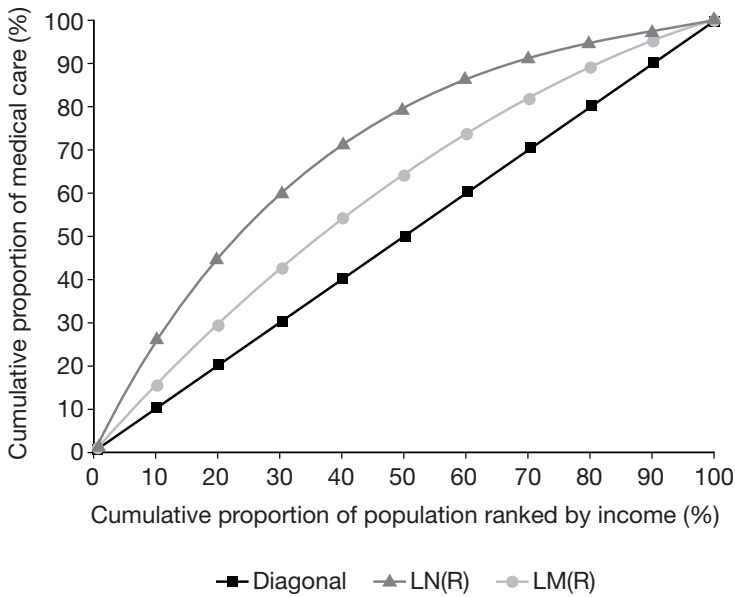

Fig. 2.6.1 Concentration curves for utilization (LM) and need (LN) compared to line of perfect equality (diagonal)

groups, implying pro-poor (pro-rich) inequality. The level of horizontal inequity in the receipt of health care is quantified by comparing the two distributions – when the unadjusted health care utilization and needs-adjusted utilization curves coincide, the horizontal inequity index equals zero (no inequity). Horizontal inequity favours the richer (poorer) if the needs-adjusted concentration curve lies above (below) the unadjusted utilization concentration curve.

Kakwani et al. have shown that it is possible to compute the index using a convenient regression of the concentration index on the relative income rank (Kakwani et al. 1997; O'Donnell et al. 2008). Based on an initial health-care demand model (as in the regression approach described above) it is possible to calculate the concentration index of needs-predicted utilization. This is compared with the concentration index of actual utilization to calculate the index of horizontal inequity.

The concentration index is therefore a relative measure of inequality (Wagstaff et al. 1989) that has the main advantages of capturing the socio-economic dimension of inequities; including information on the whole socio-economic distribution (i.e. income distribution); providing visual representation through the concentration curves; and, finally, allowing checks of stochastic relationships (Wagstaff et al. 1991). Moreover, this approach allows comparisons of inequity across countries and across time in order to understand the specific role that health system characteristics play in inequity.

Horizontal inequity indices were defined primarily to synthesize information from cross-sectional data but they have also been used to measure socio-economic inequalities in health and health-care use with longitudinal data (Bago d'Uva et al. 2007; Hernández-Quevedo et al. 2006). A longitudinal perspective enables the researcher to reveal whether inequalities have reduced or increased with time and to classify them as either short-term (using cross-sectional data) or long-term (aggregated over a series of periods) (Jones & López-Nicolás 2004). A mobility index (MI) can be created to summarize the discrepancy between short- and long-term inequalities. This is equal to one minus the ratio of the long-term inequity index and the weighted sum of all the short-term (cross-sectional) inequity indices. If the long-term index is equal to the weighted sum of the short-term inequity indices then MI equals zero. If it is negative (positive) the long-term inequity is larger (smaller) than the weighted sum of short term inequity:

$$MI = 1 - (HI^{LT}/SHI^{ST})$$

This methodology has been used mainly for analyses of inequalities in health (Hernández-Quevedo et al. 2006; Lecluyse 2007).

The concentration index approach has a further advantage of enabling decomposition of the contribution of need (i.e. ill-health) and non-need (i.e. socio-economic) variables to overall inequality in health care (O'Donnell et al. 2008; Wagstaff et al. 2003). The contribution of each determinant to total inequality in health-care utilization can be decomposed into two deterministic components (equal to the weighted sum of the concentration indices of need and non-need regressors) and a residual component that reflects the inequality in health that cannot be explained by systematic variation across income groups. Therefore, the contributors to inequality can be divided into inequalities in each of the need and non-need variables. Each variable's contribution to total inequality would be the sum of three factors: (i) the relative weight of such a variable (measured by its mean); (ii) its income distribution (indicated by the concentration index of the variable of interest); and (iii) its marginal effect on the utilization of health care (regression coefficient). Hence the decomposition method can be a useful instrument for describing the factors that contribute to inequality .

Despite the extensive use of the Concentration Index (CI), the shortcomings associated with this measure have been recently discussed in the literature. Firstly, the CI depends on the mean of the variable and, hence, could confound comparisons of health inequality across time or countries (Wagstaff 2005). Secondly, the ranking differs depending on whether one measures inequalities in health or inequalities in ill-health (Clarke et al. 2000). Finally, the value provided by the CI is arbitrary if one analyses a qualitative measure of health (Erreygers 2006). To overcome these limitations, Erreygers (2009) recently proposed a corrected version of the CI that transforms the standard index by the mean and the bounds of the health variable. This adjusted CI has already been applied in different works (for example, van de Poel et al. 2008).

The concentration index approach has been used mainly for measuring horizontal inequity – equal utilization for people with equal need, independent of income. Few studies have used the vertical equity principle of proportional unequal access for unequals. In contrast, the

vertical equity principle has been used mainly for measuring income-related equity in health-care finance (O'Donnell et al. 2008; Wagstaff & van Doorslaer 2000; Wagstaff et al. 1999). The Kakwani index measures the extent to which each source of finance (e.g. taxes, social insurance, private insurance, out-of-pocket payments) or the overall financing system (weighted average of each source of finance index) departs from proportionality.

The empirical research on equity of access to health care has increasingly drawn on the technical methods of the concentration and horizontal inequity indices (Allin et al. 2009; Chen & Escarce 2004; Jiménez-Rubio et al. 2008; Lu et al. 2007; Masseria et al. 2009; van Doorslaer et al. 2004; van Doorslaer et al. 2006). A recent OECD project evaluated income-related inequity across twenty-one countries in physician, hospital and dental sectors (van Doorslaer et al. 2004a; van Doorslaer et al. 2006), standardizing for needs (measured as self-reported health status, health limitations, age and gender). The decomposition approach was also used to disentangle the role of different need and non-need variables. The detailed results of equity in physician visits are discussed here.

Within-country variations in use by income indicate that low-income groups are more likely to visit a doctor than higher income groups in all OECD countries. However, standardizing for population needs, the probability of a doctor visit was higher among richer groups (Fig. 2.6.2). The probability of contacting a general practitioner appeared to be distributed according to need and no statistically significant inequities were found, except in Canada, Finland and Portugal. However, when considering only those who have at least one general practitioner visit, poorer people consulted general practitioners more often. The pattern was very different for specialist visits. In all countries, higher-income individuals had a significantly higher probability of visiting a specialist, and were making more visits, than the poor.

The authors followed the decomposition method to calculate the contributions of need, income, education, activity status, region and insurance to total inequality. Fig. 2.6.3 reports the results for the analysis of specialist visit probability. The contribution of need was negative in all countries (it reduced inequity) but the contribution of income, education and insurance was positive. Table 2.6.3 examines the role of education in inequity in the probability of a specialist visit in Spain.

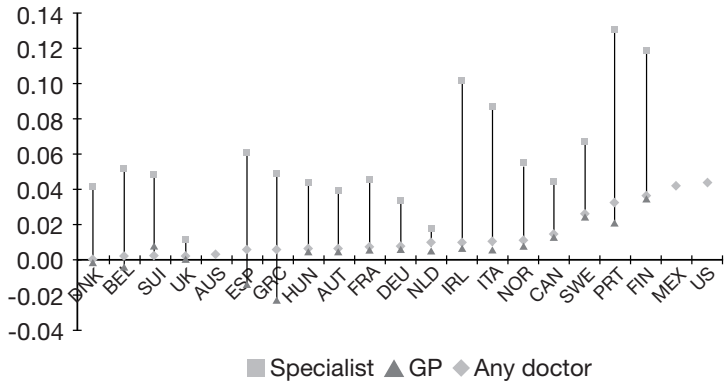

Fig. 2.6.2 Horizontal inequity indices for annual probability of a visit, twenty-one OECD countries

Countries ranked by HI index for doctor visits. HI indices are estimated as concentration indices for need-standardized use. Positive (negative) index indicates a pro-rich (pro-poor) distribution. German general practitioner and specialist indices calculated from ECHP 1996

Source: van Doorslaer, Masseria & Koolman 2006

Low education's contribution to inequity depends on its mean value (63% of the population reported to have low education); relationship with income (measured by the concentration index which indicates that people with low education tend also to have lower incomes); and marginal effect on specialist care (people with low education use specialist care 4.3% less than those with higher education). Thus poor education makes a positive contribution to total inequality, thereby increasing inequity. The total contribution of education is given by the sum of the contributions of low and medium education.

A longitudinal perspective enables the researcher to reveal whether inequalities have reduced or increased with time. Hospital care is a particularly interesting example of the usefulness of this data. Infrequent annual use of hospital care and its skewed distribution may undermine the reliability of estimates of hospital care needs in cross-sectional analysis, particularly when the sample size is relatively small. Masseria et al. (2009) compared the pooled (1994-1998) and wave by wave results of the ECHP. They demonstrated that it was possible to enhance the power of the estimates and to obtain robust estimates of inpatient horizontal inequity by pooling several years of survey data, (see Table 2.6.4). Indeed, inequity in hospital care was found to be

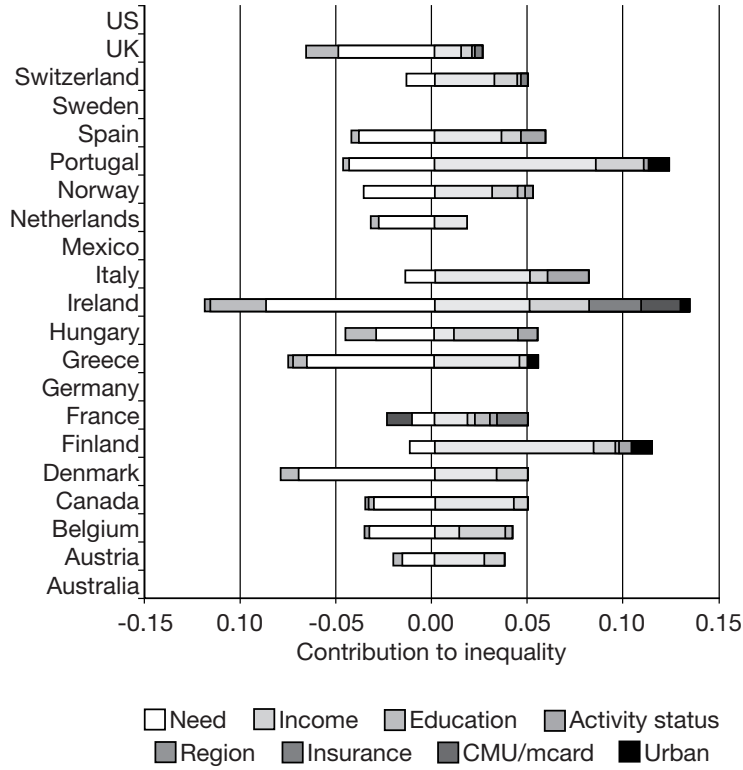

Fig. 2.6.3 Decomposition of inequity in specialist probability

Source: van Doorslaer et al. 2004a

Table 2.6.3 ***Contribution of income and education to total specialist inequality in Spain, 2000***

	Mean	Concentration index	Marginal effect	Contribution to inequity	Sum contribution
HI index				0.066	0.066
Logarithm of income	14.121	0.025	0.047	0.036	0.036
Education: medium	0.171	0.139	-0.008	0.000	
Education: low	0.630	-0.159	-0.043	0.010	0.009

Source: van Doorslaer et al. 2004a

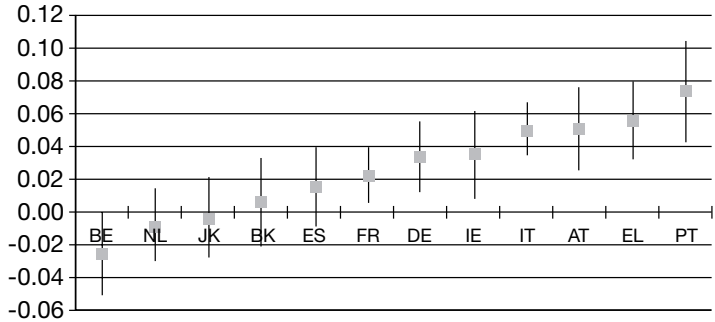

Fig. 2.6.4 Horizontal inequity index for the probability of hospital admission in twelve European countries (1994-1998)

Source: Masseria et al. 2009

significantly pro-rich in seven of the twelve countries analysed and significantly pro-poor in one – Belgium. Conversely, the wave by wave results rarely showed significant inequity, due to their lack of power.

In Table 2.6.4, the MI summarizes the discrepancy between short- and long-term inequalities. The MI was found to be negative in some countries and positive in others. Negative mobility indices mean that the weighted averages of the cross-sectional concentration indices are smaller in absolute value than the longitudinal indices. A negative index suggests that individuals with downwardly mobile incomes have below average levels of health-care use compared to upwardly mobile individuals. This makes long-run income-related inequity greater than would be expected from a cross-sectional measure (contrary applies to a positive index).

Policy implications and directions for future research

Most governments widely accept the goal of equitable access to health care. This goal is motivated by the egalitarian view that access to care is a right and by the potential for equity of access to help reduce health inequalities. Translating this policy goal to a measurable objective is not straightforward. Moreover, considerable debate surrounds the definition of equity, health-care need and access as well as the methods for calculating equity in health care.

Empirical research most commonly measures the goal of treating equals equally; health-care need is measured by levels of ill-health and

Table 2.6.4 ***Short-run and long-run horizontal inequity index, MI***

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Pooled	Mobility
Austria		**0.046**	**0.070**	**0.052**	0.036	**0.050**	0.029
Belgium	-0.04	-0.029	0.003	-0.019	-0.046	**-0.025**	-0.031
Denmark	0.00	0.049	-0.022	0.022	-0.022	0.006	-0.120
France	0.01	-0.011	0.026	0.030	**0.075**	**0.023**	0.085
Germany	**0.03**	**0.056**	0.015			**0.033**	0.005
Greece	**0.07**	**0.060**	0.037	0.031	**0.074**	**0.055**	-0.015
Ireland	0.04	0.039	0.077	-0.017	0.050	**0.036**	0.025
Italy	0.02	**0.066**	**0.059**	**0.040**	**0.067**	**0.050**	-0.056
Netherlands	0.02	**-0.049**	-0.009	0.029	-0.024	-0.008	0.058
Portugal	0.04	**0.071**	0.087	0.100	0.082	0.074	-0.082
Spain	0.03	0.000	**0.041**	-0.026	0.037	0.016	-0.032
UK	0.00	-0.010	-0.001			-0.003	0.193

Numbers in bold are statistically significant with a 95% confidence interval
Source: authors' calculations based on Masseria et al. 2009

access approximated by utilization. Thus, inequity can be identified where patterns of utilization differ between individuals with the same health-care need (health status and risk of ill-health) across income, social or other socio-economic groups. These analyses require information on socio-economic status, health status and utilization patterns, whether using regression methods or calculating concentration indices of inequity. Analyses of equity can be used to inform policy decisions insofar as the studies are based on accurate and meaningful data.

Empirical analyses may be based on survey, administrative or, ideally, linked datasets. Survey data provide comprehensive information on all these levels but administrative data may provide more accurate information on utilization. This can include the intensity of use measured not just by number of visits but also by total expenditure and the different types of services used (e.g. diagnostic tests received, day surgeries, referrals). Administrative utilization data also address the problems of recall bias and subjectivity, and cover the entire population using health care including those groups typically excluded or

underrepresented in surveys (people who are homeless, without telephones or living in institutions). However, administrative data provide a less comprehensive source of socio-economic information and health status. Socio-economic data would typically be collected through geographical measures of income or deprivation. Health status could be measured by physician diagnosis but this limits the information available to those who have been in contact with the health system. Linking administrative and survey data is the ideal approach to benefit from the accuracy and detail of utilization information and the comprehensiveness of self-reported socio-economic and health indicators from surveys.

The majority of studies draw on survey data to undertake equity analyses. Self-reported indicators of health status are the most commonly used measures of health-care needs as they are available in national and international health surveys. These measures are subject to numerous methodological problems but various studies have shown that they are strong predictors of objective health status and mortality. However, even if ill-health is measured accurately it may not provide an indication of what (and to what extent) services are needed to restore health (Culyer & Wagstaff 1993). A review of equity studies in the United Kingdom noted that the majority pay little attention to the complex concept of need (Goddard & Smith 2001). The majority of studies show widespread acceptance of the assumption that need can be measured using SAH, though many also control for factors that may affect the reporting of health status (e.g. age and sex) and incorporate some indication of an individual's risk of ill-health (e.g. age, obesity, symptoms), while also considering a broad set of SAH indicators.

There has been some growth in the collection of more objective indicators of health. Recent health surveys (e.g. SHARE, Health Interview Survey) include quasi-objective indicators of ill-health, based on respondents' reporting on more factual items such as specific conditions or activity limitations (e.g. presence of chronic conditions, specific limitations in ADL or IADL). These indicators have proved useful for building a more general index of ill-health that corrects issues of reporting bias (Jürges 2007). A few surveys (e.g. WHS) have recently introduced vignettes that allow potential biases to be corrected with SAH measures. The availability of objective measures of health, such as biomarkers, is restricted to a few national, cross-sectional surveys

and still presents a methodological issue concerning the standardization of data collection.

The methodological difficulties associated with measuring equity are discussed above. In addition, needs-adjusted utilization does not account for potentially acceptable variations in utilization, such as those driven by individuals' choices (Le Grand 1991; Whitehead 1991). Survey data permit further subjective analyses of health-care contacts such as perceived timeliness, quality and overall satisfaction that complement information on utilization. Moreover, subjective unmet need for health care may also be included in surveys. Subjective unmet need has largely been interpreted to represent system-level barriers to access (Elofsson et al. 1998; Mielck et al. 2007; Westin et al. 2004). However, the different reasons for unmet need include personal (e.g. fears and preferences) and system factors (e.g. costs). It is important to differentiate these reasons and to examine the association between reported unmet need and contacts with the health system. Research linking information on levels and reasons for subjective unmet need with actual health-care utilization patterns could therefore complement conventional equity analyses.

Meaningful research on equity in health care relies on the availability of comprehensive and reliable data. Ideally, these would be longitudinal survey and administrative sources linked at the individual level. Population health surveys should include information on health status (including general, specific, subjective and quasi-objective measures, vignettes to test for reporting bias); socio-economic status (including all income sources, assets such as home ownership and financial assets, education, employment); utilization of health care (disaggregated by type of service); experiences with health care (including accessibility, acceptability, waiting times, satisfaction, perceived quality, direct costs, non-use of health care, i.e. unmet need); and other factors that affect access (including details of insurance status and entitlements). Furthermore, information on an individual's residence (post code) makes it possible to calculate the distance to health-care facilities. Finally, clinical appropriateness could be assessed on the basis of available information on diagnoses and health service utilization. This quality aspect of health care remains relatively undeveloped in equity analyses.

Longitudinal data permit more in-depth investigation of the trends and dynamics of inequalities over time. The long-term perspective

provides useful information on population-representative disease trajectories; links between outcomes and earlier experiences and behaviours; and the dynamics between individual and family characteristics, take-up of insurance, asset accumulation, health and health care. For the measurement of inequalities in health, it has been shown that the use of longitudinal data captures the mobility of individuals in their ranking according to their socio-economic levels (Hernández-Quevedo et al. 2006; Jones & López-Nicolás 2004). Such mobility is particularly interesting if this variation is systematically associated with changes in levels of health (Hernández-Quevedo et al. 2006). For the study of equity of access to health care, longitudinal data also allow consideration of the possible endogeneity of need variables in the health-care utilization models (Sutton et al. 1999).

A growing evidence base demonstrates inequitable utilization or treatment patterns in many countries, though many questions remain (including whether inequity of access to health care contributes to inequalities in health). There is a need to investigate the link between access to health care, health outcomes and health inequalities. This will not only improve understanding of the processes by which health inequalities arise and can be reduced, but also may increase support for improving efforts to ensure equitable access. It is difficult to address the question of whether inequitable utilization leads to unequal health outcomes on a population level. The research that has been conducted has relied on disease-specific approaches which (although not generalizable to the population level) have the potential to inform policy approaches, e.g. in the treatment of particular conditions such as acute myocardial infarction in Canada (Alter et al. 1999; Alter et al. 2006; Pilote et al. 2003).

It is well-known that the policies needed to reduce inequalities in health call for integrated, multi-sectoral approaches that extend beyond the health system (Mackenbach & Bakker 2002; WHO 2008). These address not only health and social care and poverty alleviation but also health-related behaviours (smoking, alcohol consumption, diet, obesity); psychosocial factors (psychosocial stressors, social support, social integration); material factors (housing conditions, working conditions, financial problems); and access to health care. Many countries have explicit public health policies that address some or all of these (Judge et al. 2006). Equitable access to health care plays a critical role (Dahlgren & Whitehead 2006). Careful monitoring of equity in health

care on the basis of robust empirical analyses is vital to measure the impact of health-care policies and broader reform initiatives on health system performance. Continued research is needed to understand not only the causes of inequity but also what policy measures are effective in ensuring that individuals in need receive effective, high-quality health care.

References

Abásolo, I. Manning, R. Jones, A (2001). 'Equity in utilization of and access to public-sector GPs in Spain.' *Applied Economics,* 33(3): 349–364.

Adams, P. Hurd, M. McFadden, D. Merrill, A. Ribeiro, T (2003). 'Healthy, wealthy and wise? Tests for direct causal paths between health and socioeconomic status.' *Journal of Econometrics,* 112: 3–56.

Aday, LA. Andersen, RM (1974). 'A framework for the study of access to medical care.' *Health Services Research,* 9(3): 208–220.

Aday, LA. Andersen, RM (1981). 'Equity of access to medical care: a conceptual and empirical overview.' *Medical Care,* 19(12): 4–27.

Allin, S. Masseria, C. Mossialos, E (2009). 'Measuring socioeconomic differences in use of health care services by wealth versus by income.' *American Journal of Public Health*, 10.2105/AJPH.2008.141499.

Alter, DA. Chong, A. Austin, PC. Mustard, C. Iron, K. Williams, JI. Morgan, CD. Tu, JV. Irvine, J. Naylor, CD. SESAMI Study Group (2006). 'Socioeconomic status and mortality after acute myocardial infarction.' *Annals of Internal Medicine,* 144(2): 82–93.

Alter, DA. Naylor, DC. Austin, P. Tu, JV (1999). 'Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. *New England Journal of Medicine,* 341(18): 1359–1367.

Andersen, RM (1995). 'Revisiting the behavioral model and access to medical care: does it matter?' *Journal of Health and Social Behaviour,* 36(1): 1–10.

Angel, R. and Thoits, P (1987). 'The impact of culture on the cognitive structure of illness.' *Culture, Medicine and Psychiatry,* 11(4): 465–494.

Atkinson, A. Cantillon, B. Marlier, E. Nolan, B (eds.) (2002). '*Social indicators: the EU and social inclusion.*' Oxford: Oxford University Press.

Bago d'Uva, T. Jones, A. van Doorslaer, E (2007). *Measurement of horizontal inequity in health care utilization using European panel data.* Rotterdam: Erasmus University (Tinbergen Institute Discussion Paper TI 2007 - 059/3).

Bago d'Uva, T. Van Doorslaer, E. Lindeboom, M. O'Donnell, O (2008). 'Does reporting heterogeneity bias the measurement of health disparities?' *Health Economics,* 17(3): 351–375.

Bailis, DS. Segall, A. Chipperfield, JG (2003). 'Two views of self-rated general health status.' *Social Science and Medicine,* 56(2): 203–217.

Banks, J. Marmot, M. Oldfield, Z. Smith, JP (2006). 'Disease and disadvantage in the United States and England.' *Journal of the American Medical Association,* 295(17): 2037–2045.

Barsky, AJ. Cleary, PD. Klerman, GL (1992). 'Determinants of perceived health status of medical outpatients.' *Social Science and Medicine,* 34(10): 1147–1154.

Bevan, G (2008). *Review of the weighted capitation formula.* London: Department of Health.

Buchmueller, T. Grumbach, K. Kronick, R. Kahn, JG (2005). 'The effect of health insurance on medical care utilization and implications for insurance expansion: a review of the literature.' *Medical Research and Review,* 62(1): 3–30.

Chen, AY. Escarce, JJ (2004). 'Quantifying income-related inequality in healthcare delivery in the United States.' *Medical Care,* 42(1): 38–47.

Clarke, PM. Gerdtham, U-G. Johannesson, M. Bingefors, K. Smith, L (2002). On the measurement of relative and absolute income-related health inequality. *Social Science & Medicine*, 55(11):1923–1928.

Collins, E. Klein, R (1980). 'Equity and the NHS: self-reported morbidity, access and primary care.' *British Medical Journal,* 281(6248): 1111–1115.

Crossley, TF. Kennedy, S (2002). 'The reliability of self-assessed health status.' *Journal of Health Economics,* 21(4): 643–658.

Culyer, AJ (1995). 'Need: the idea won't do – but we still need it.' *Social Science and Medicine,* 40(6): 727–730.

Culyer, AJ (2007). 'Equity of what in healthcare? Why the traditional answers don't help policy – and what to do in the future?' *Healthcare Papers,* 8(Spec. No.): 12–26.

Culyer, AJ. Wagstaff, A (1993). 'Equity and equality in health and health care.' *Journal of Health Economics,* 12(4): 431–457.

Dahlgren, G. Whitehead, M (2006). *European strategies for tackling social inequities in health: levelling up. Part 2.* Copenhagen: WHO Regional Office for Europe.

Department of Health (2002). *Tackling inequalities in health: 2002 cross-cutting review.* London: The Stationery Office.

Department of Health (2003). *Tackling inequalities in health: a programme for action.* London: The Stationery Office.

Donabedian, A (1971). 'Social responsibility for personal health services: an examination of basic values.' *Inquiry,* 8(2): 3–19.

Donabedian, A (1972). 'Models for organizing the delivery of personal health services and criteria for evaluating them.' *Milbank Memorial Fund Quarterly*, 50(Pt 2): 103–154.

Dunlop, PC. Coyte, PC. McIsaac, W (2000). 'Socio-economic status and the utilisation of physicians' services: results from the Canadian National Population Health Survey.' *Social Science and Medicine,* 51(1): 123–133.

Elofsson, S. Undén, A-L. Krakau, I (1998). 'Patient charges – a hindrance to financially and psychosocially disadvantaged groups seeking care.' *Social Science and Medicine,* 46(10): 1375–1380.

Erreygers, G (2006). Beyond the health Concentration Index: an Atkinson alternative for the measurement of the socioeconomic inequality of health. In: Paper presented at the Conference Advancing Health Equity, Helsinki, WIDER-UNU.

Erreygers, G (2009). Correcting the Concentration Index. *Journal of Health Economics*, 28(2): 504–515.

Evans, RG (1994). Introduction. In: Evans, RG. Marmor, T. Barer, M (eds.). *Why are some people healthy and others not? The determinants of health of populations*. Berlin: Aldine de Gruyter.

Folland, S. Goodman, AC. Stano, M (2004). *The economics of health and health care*. Upper Saddle River, NJ: Pearson Prentice Hall.

Goddard, M. Smith, P (2001). 'Equity of access to health care services: theory and evidence from the UK.' *Social Science and Medicine,* 53(9): 1149–1162.

Gravelle, H. Morris, S. d Sutton, M (2006). Economic studies of equity in the consumption of health care. In: Jones, AJ (ed.). *The Elgar companion to health economics*. Cheltenham: Edward Elgar.

Groot, W (2000). 'Adaptation and scale of reference bias in self-assessments of quality of life.' *Journal of Health Economics,* 19(3): 403–420.

Grossman, M (1972). 'On the concept of health capital and the demand for health.' *Journal of Political Economy,* 80(2): 223–255.

Gulliford, M (2002). Equity and access to health care. In: Gulliford, M. Morgan, M (eds.). *Access to health care*. London: Routledge.

Gulliford, M. Figueroa-Minoz, J. Morgan, M (2002a). Meaning of 'access' in health care. In: Gulliford, M. Morgan, M (eds.). *Access to health care*. London: Routledge.

Gulliford, M. Figueroa-Munoz, J. Morgan, M. Hughes, D. Gibson, B. Beech, R. Hudson, M (2002b). 'What does 'access to health care' mean?' *Journal of Health Services Research and Policy,* 7(3): 186–188.

Häkkinen, U. Luoma, K (2002). 'Change in determinants of use of physician services in Finland between 1987 and 1996.' *Social Science and Medicine,* 55(9): 1523–1537.

Healy, J. McKee, M (eds.) (2004). *Accessing health care: responding to diversity.* Oxford: Oxford University Press.

Hernández-Quevedo, C. Jones, A. López-Nicolás, A. Rice, N (2006). 'Socioeconomic inequalities in health: a comparative longitudinal analysis using the European Community household panel.' *Social Science and Medicine,* 63(5): 1246–1261.

Hernández-Quevedo, C. Jones, A. Rice, N (2008). 'Reporting bias and heterogeneity in self-assessed health. Evidence from the British Household Panel Survey.' [in Spanish] *Cuadernos Económicos de ICE,* 75: 63–97.

Hurley, J (2000). An overview of the normative economics of the health sector. In: Culyer, AJ. Newhouse, JP (eds.). *Handbook of health economics*. Amsterdam: Elsevier Science BV.

Hurst, J. Jee-Hughes, M (2001). Performance measurement and performance management in OECD health systems. In: *Labour market and social policy occasional papers no. 47*. Paris: OECD.

Idler, E. Benyamini, Y (1997). 'Self-rated health and mortality: a review of twenty-seven community studies.' *Journal of Health and Social Behavior,* 38(1): 21–37

Idler, E. Kasl, SV (1995). 'Self-ratings of health: do they also predict change in functional ability?' *Journal of Gerontology,* 50(6): 344–353.

Jiménez-Rubio, D. Smith, PC. van Doorslaer, E (2008). 'Equity in health and health care in a decentralised context: evidence from Canada.' *Health Economics,* 17(3): 377–392.

Johnston, DW. Propper, C. Shields, MA (2007). *Comparing subjective and objective measures of health: evidence from hypertension for the income/health gradient*. Bonn: Institute for the Study of Labor (IZA Discussion Paper No. 2737).

Jones, AJ. López-Nicolás, A (2004). 'Measurement and explanation of socioeconomic inequality in health with longitudinal data.' *Health Economics,* 13(10): 1015–1030.

Judge, K. Platt, S. Costongs, C. Jurczak, K (2006). Health inequalities: a challenge for Europe. In: *Report prepared for the UK Presidency of the EU*. London: Department of Health.

Jürges, H (2007). 'True health vs response styles: exploring cross-country differences in self-reported health.' *Health Economics,* 16(2): 163–178.

Kakwani, N. Wagstaff, A. van Doorslaer, E (1997). 'Socioeconomic inequality in health: measurement, computation and statistical inference.' *Journal of Econometrics,* 77(1): 87–103.

Kenkel, D (1995). 'Should you eat breakfast? Estimates from health production functions.' *Health Economics,* 4(1): 15–29.

Kerkhofs, M. Lindeboom, M (1995). 'Subjective health measures and state dependent reporting errors.' *Health Economics,* 4(3): 221–235.

Koolman, X. van Doorslaer, E (2004). 'On the interpretation of the concentration index of inequality.' *Health Economics,* 13(7): 649–656.

Lecluyse, A (2007). 'Income-related health inequality in Belgium: a longitudinal perspective.' *European Journal of Health Economics*, 8(3): 237–243.

Le Grand, J (1978). 'The distribution of public expenditure: the case of health care.' *Economica,* 45(178): 125–142.

Le Grand, J (1982). *The strategy of equality.* London: George Allen and Unwin.

Le Grand, J (1991). *Equity and choice: an essay in economics and applied philosophy.* London: Harper Collins Academic.

Lindeboom, M. van Doorslaer, E (2004). 'Cut-point shift and index shift in self-reported health.' *Journal of Health Economics,* 23(6): 1083–1099.

Lu, J. F. Leung, GM. Kwon, S. Tin, KY. van Doorslaer, E. O'Donnell, O (2007). 'Horizontal equity in health care utilization evidence from three high-income Asian economies.' *Social Science and Medicine,* 64(1): 199–212.

Mackenbach, JP. Bakker, MJ (2002). *Reducing inequalities in health.* London: Routledge.

Mackenbach, JP. Kunst, AE (1997). 'Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe.' *Social Science and Medicine,* 44(6): 757–771.

Mackenbach, JP. Simon, JG. Looman, CWN. Joung, IMA (2002). 'Self-assessed health and mortality: could psychosocial factors explain the association?' *International Journal of Epidemiology,* 31(6): 1162–1168.

Masseria, C. Allin, S. Sorenson, C. Papanicolas, I. Mossialos, E (2007). *What are the methodological issues related to measuring health and drawing comparisons across countries? A research note*. Brussels: DG Employment and Social Affairs, European Observatory on the Social Situation and Demography.

Masseria, C. Koolman, X. van Doorslaer, E (2009). 'Income related inequality in the probability of a hospital admission in Europe.' *Health Economics Policy and Law,* forthcoming.

McGee, DL. Liao, Y. Cao, G. Cooper, RS (1999). 'Self-reported health status and mortality in a multiethnic US cohort.' *American Journal of Epidemiology,* 149(1): 41–46.

Mielck, A. Kiess, R. van den Knesebeck, O. Stirbu, I. Kunst, A (2007). Association between access to health care and household income among the elderly in 10 western European countries. In: *Tackling health inequalities in Europe: an integrated approach*. Rotterdam: Erasmus MC Department of Public Health.

Mooney, G (1983). 'Equity in health care: confronting the confusion.' *Effective Health Care,* 1(4): 179–185.

Mooney, G (1986). *Economics, medicine and health care*. Brighton: Wheatsheaf Books Ltd.

Morris, S. Sutton, M. Gravelle, H (2005). 'Inequity and inequality in the use of health care in England: an empirical investigation.' *Social Science and Medicine,* 60(6): 1251–1266.

Mossey, J. Shapiro, E (1982). 'Self-rated health: a predictor of mortality among the elderly.' *American Journal of Public Health,* 72(8): 800–808.

Mossialos, E. Thomson, S (2003). 'Access to health care in the European Union: the impact of user charges and voluntary health insurance. In: Gulliford, M. Morgan, M (eds.) *Access to health care.* London: Routledge.

O'Donnell, O. van Doorslaer, E. Wagstaff, A. Lindelow, M (2008). *Analyzing health equity using household survey data: a guide to techniques and their implementation.* Washington, DC: The World Bank.

OECD (1992). *The reform of health care: a comparative analysis of seven OECD countries*. Paris.

Oliver, A. Mossialos, E (2004). 'Equity of access to health care: outlining the foundation for action.' *Journal of Epidemiology and Community Health,* 58(8): 655–658.

Pilote, L. Joseph, L. Bélisle, P. Penrod, J (2003). 'Universal health insurance coverage does not eliminate inequities in access to cardiac procedures after acute myocardial infarction.' *American Heart Journal,* 146(6): 1030–1037.

President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioural Research (1983). '*Securing access to health care*.' Washington, DC: US Government Printing Office.

Rawls, J (1971). *A theory of justice.* Cambridge, Massachusetts: Harvard University Press.

Sadana, R. Mathers, CD. Lopez, AD. Murray, CJL. Iburg, K (2000). *Comparative analysis of more than 50 household surveys on health status*. Geneva: World Health Organization (GPE Discussion Paper No 15. EIP/GPE/EBD).

Sen, A (1992). *Inequality reexamined.* Cambridge: Harvard University Press.

Sen, A (2002). 'Health: perception versus observation.' *British Medical Journal,* 324(7342): 860–861.

Singh-Manoux, A. Martikainen, P. Ferrie, J. Zins, M. Marmot, M. Goldberg, M (2006). 'What does self rated health measure? Results from the British Whitehall II and French Gazel cohort studies.' *Journal of Epidemiology and Community Health,* 60(4): 364–372.

Starfield, B (1993). *Primary care – concept, evaluation and policy.* Oxford: Oxford University Press.

Sundquist, J. Johansson, SE (1997). 'Self-reported poor health and low educational level predictors for mortality: a population-based follow

up study of 39,156 people in Sweden.' *Journal of Epidemiology and Community Health,* 51(1): 35–40.

Sutton, M. Carr-Hill, R. Gravelle, H. Rice, N (1999). 'Do measures of self-reported morbidity bias the estimation of the determinants of health care utilization?' *Social Science and Medicine,* 49(7): 867–878.

Thiede, M. Akweongo, P. McIntyre, D (2007). Exploring the dimensions of access. In: McIntyre, D. Mooney, G (eds.). *The economics of health equity.* Cambridge: Cambridge University Press.

Van de Poel, E. Hosseinpoor, A, Speybroeck, N. Van Ourti, T. Vega, J (2008). Socioeconomic inequality in malnutrition in developing countries. Bulletin of the World Health Organization, 84(4): 282–291.

Van der Heyden, JH. Demarest, S. Tafforeau, J. Van Oyen, H (2003). 'Socio-economic differences in the utilization of health services in Belgium.' *Health Policy,* 65(2): 153–165.

van Doorslaer, E. Gerdtham, UG. (2003). 'Does inequality in self-assessed health predict inequality in survival by income? Evidence from Swedish data.' *Social Science and Medicine,* 57(9): 1621–1629.

van Doorslaer, E. Koolman, X. Jones, A (2004). 'Explaining income-related inequalities in doctor utilisation in Europe.' *Health Economics,* 13(7): 629–647.

van Doorslaer, E. Masseria, C. Koolman, X. OECD Health Equity Research Group (2006). 'Inequalities in access to medical care by income in developed countries.' *Canadian Medical Association Journal,* 174(2): 177–183.

van Doorslaer, E. Masseria, C. OECD Health Equity Research Group Members (2004a). *Income-related inequality in the use of medical care in 21 OECD countries*. Paris: OECD.

van Doorslaer, E. Wagstaff, A. Bleichrodt, H. Calonge, S. Gerdtham, UG. Gerfin, M. Geurts, J. Gross, L. Häkkinen, U. Leu, RE. O'Donnell, O. Propper, C. Puffer, F. Rodríguez, M. Sundberg, G. Winkelhake, O (1997). 'Income-related inequalities in health: some international comparisons.' *Health Economics,* 16(1): 93–112.

van Doorslaer, E. Wagstaff, A. Rutten, F (eds.) (1993). *Equity in the finance and delivery of health care: an international perspective.* Oxford: Oxford University Press.

van Doorslaer, E. Wagstaff, A. van der Burg, H. Christiansen, T. De Graeve, D. Duchesne, I. Gerdtham, UG. Gerfin, M. Geurts, J. Gross, L. Hakkinen, U. John, J. Klavus, J. Leu, RE. Nolan, B. O'Donnell, O. Propper, C. Puffer, F. Schellhorn, M. Sundberg, G. Winkelhake, O (2000). 'Equity in the delivery of health care in Europe and the US.' *Journal of Health Economics,* 19(5): 553–583.

Wagstaff, A (2005). The bounds of the Concentration Index when the variable of interest is binary, with an application to immunization inequality. *Health Economics*, 14(4): 429–432.

Wagstaff, A. Paci, P. van Doorslaer, E (1989). 'Equity in the finance and delivery of health care: some tentative cross-country comparisons.' *Oxford Review of Economic Policy,* 5(1): 89–112.

Wagstaff, A. van Doorslaer, E (2000). Equity in health care finance and delivery. In: Culyer, AJ. Newhouse, JP (eds.). *Handbook of health economics*. Amsterdam: North-Holland, pp. 1803–1862.

Wagstaff, A. van Doorslaer, E. Paci, P (1991). 'On the measurement of horizontal inequity in the delivery of health care.' *Journal of Health Economics,* 10(2): 169–205.

Wagstaff, A. van Doorslaer, E. van Der Burg, H. Calonge, S. Christiansen, T. Citoni, G. Gerdtham, UG. Gerfin, M. Gross, L. Häkinnen, U. Johnson, P. John, J. Klavus, J. Lachaud, C. Lauritsen, J. Leu, R. Nolan, B. Perán, E. Pereira, J. Propper, C. Puffer, F. Rochaix, L. Rodríguez, M. Schellhorn, M. Winkelhake, O. et al (1999). 'Equity in the finance of health care: some further international comparisons.' *Journal of Health Economics,* 18(3): 263–290.

Wagstaff, A. van Doorslaer, E. Watanabe, N (2003). 'On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam.' *Journal of Econometrics,* 112(1): 207–223.

Westin, M. Ahs, A. Persson, KB. Westerling, R (2004). 'A large proportion of Swedish citizens refrain from seeking medical care – lack of confidence in the medical services a plausible explanation?' *Health Policy,* 68(3): 333–344.

Whitehead, M (1991). 'The concepts and principles of equity and health.' *Health Promotion International,* 6(3): 217–228.

WHO (2000). *The world health report 2000. Health systems: improving performance*. Geneva: World Health Organization.

WHO (2008). *Closing the gap in a generation. Health equity through action on the social determinants of health*. Geneva: World Health Organization.

Williams, A (1993). Equity in health care: the role of ideology. In: van Doorslaer, E. Wagstaff, A. Rutten, F (eds.). *Equity in the finance and delivery of health care*. Oxford: Oxford University Press.

Williams, A (2005). The pervasive role of ideology in the optimisation of the public-private mix in public healthcare systems. In: Maynard, A (ed.). *The public-private mix for health*. London: The Nuffield Trust.

Zimmer, Z. Natividad, J. Lin, HS. Chayovan, N (2000). 'A cross-national examination of the determinants of self-assessed health.' *Journal of Health and Social Behavior,* 41(4): 465–481.

2.7 *Health system productivity and efficiency*

ANDREW STREET, UNTO HÄKKINEN

Introduction

In the light of apparently inexorable rises in health-care expenditure, the cost effectiveness of the health system has become a dominant concern for many policy-makers. Do the funders of the health system (taxpayers, insurees, employers or patients) get good value for money? Productivity measurement is a fundamental requirement for securing providers' accountability to their payers and ensuring that health system resources are spent wisely.

Productivity measurement spans a wide range – from the cost effectiveness of individual treatments or practitioners to the productivity of a whole system. Whatever level of analysis is used, a fundamental challenge is the need to attribute both the consumption of resources (costs) and the outcomes achieved (benefits) to the organizations or individuals under scrutiny. The diverse methods used include direct measurement of the costs and benefits of treatment; complex econometric models that yield measures of comparative efficiency; and attempts to introduce health system outcomes into national accounts.

Productivity analysis can be considered via two broad questions: (i) how are resources being used? and (ii) is there scope for better utilization of these resources? These questions can be considered for the whole health system and for organizations within it but most applied research at system level tends to concentrate on the first question. The second question is the primary concern of organizational studies.

This chapter begins with an outline of the fundamental concepts required for productivity analysis, distinguishing productivity from efficiency. This is followed by a discussion of the challenges associated with applying these concepts in the health sector in which it is particularly difficult to define and measure outputs and to determine the relationship between health-care resources (inputs) and outputs.

The chapter continues with an assessment of the use of resources, as posed in the first question. Usually, the concept of productivity is of primary interest in macro-level applications, such as when considering how well an entire health system is using its resources or in analysing labour productivity over time. A growth accounting perspective is often adopted when the objective is to relate a change in outputs to a change of inputs. The productivity change of specific, common and serious health problems has also been analysed by ascribing a monetary value to outputs and relating them to the cost of treating the problem in order to evaluate value for money. In some ways, cost-effectiveness analysis which compares the benefits and cost of two or more health-care services or treatments (health technology assessment) can be seen as a form of productivity analysis. An overview of this type of approach is provided.

A range of methods have been used to consider the second question. The concept of efficiency is usually applied when considering the relative performance of organizations within a health system. These are organizations engaged in production (converting inputs into outputs) and can be hospitals, nursing homes, health centres or individual physicians. Generally speaking, such organizations face few of the competitive pressures that would encourage them to innovate and adopt cost minimizing behaviour. Comparative or benchmarking exercises aim to identify which organizations have more efficient overall operations or specific areas of operation. This information may be used to stimulate better use of resources, either by encouraging organizations to act of their own volition or through tailored incentives imposed by a regulatory authority. The final section of the chapter describes the efficiency analysis techniques that have emerged within the broad evaluative tradition.

Conceptual issues

Four fundamental questions are addressed in this section.

1. What is the relationship between inputs and outputs – i.e. what is the nature of the production process?
2. What does productivity mean and how is this concept distinct from efficiency?

3. What is the output of the health system and of the organizations within the system?
4. What resources (inputs) are employed to produce these outputs?

However, the answers are not straightforward.

Production function – relationship between inputs and outputs

The fundamental building block of productivity or efficiency analysis is the production function. This can be specified for the economy as a whole (macro-level) or for organizations within the economy (meso-level). A more technical description of the macro and meso production functions and their relationships are shown in Box 2.7.1.

Box 2.7.1 Macro-level and meso-level production functions

The production function can be applied at macro-level (for the economy as a whole) or at meso-level (for an organization within the economy). In theory, it is possible to aggregate the production functions for every organization into a function for the economy as a whole, just as total consumer spending is the sum of decisions made by many households.

The standard Cobb-Douglas production function is a useful starting point in which output (Y) is a function of two inputs – labour (L) and capital (K):

1. $$Y = AL^{\alpha}K^{\beta}$$

For calculation purposes this is transformed into logarithmic form, becoming:

2. $$\log Y = \log A + \alpha \log L + \beta \log K$$

In macro-level applications, growth accounting methods are used to assess the contribution of inputs to aggregate output growth and to estimate total productivity change for the economy as a whole or for sectors within it (Jorgenson & Griliches 1967; OECD 2001). These calculations rely on time series data, used to calculate output growth and input growth. The growth in output is defined as:

3. $$\Delta \log Y = \Delta \log A + \alpha \Delta \log L + \beta \Delta \log K$$

Where $\Delta \log Y = \log(Y_t - Y_{t-1})$; $\Delta \log L = \log(L_t - L_{t-1})$; and $\Delta \log K = \log(K_t - K_{t-1})$ with t indexing time. The parameters α and β are usually calculated as the share of income attributable to each input. The fundamental purpose of the growth accounting method is to calculate ΔA which measures the growth in output over and above the growth in inputs. This is termed total factor productivity and, when positive, is interpreted as being due to improvements in methods of production or technical progress. This interpretation rests on three key assumptions: (i) competitive factor markets; (ii) full input utilization; and (iii) constant returns to scale, $\alpha + \beta = 1$ (Inklaar et al. 2005).

Meso-level applications allow analysts to relax assumptions of constant returns to scale and to estimate more flexible functional forms than the Cobb-Douglas. Such applications use organizational data to estimate the production function from observed behaviour, either at a single time point (cross-sectional analysis) or over several time periods (panel data analysis). With cross-sectional data for a set of organizations the Cobb-Douglas production function is estimated as:

4. $$y_i = A + \hat{\alpha} \log L_i + \hat{\beta} \log K_i + \hat{\varepsilon}_i$$

Where y_i is the observed output for organization i, $i = 1...I$; L_i and K_i measure labour and capital input use for organization i; A is an estimated constant; and $\hat{\varepsilon}_i$ is the residual. The purpose is to estimate the relationships between labour and capital and output, given by the estimated parameters $\hat{\alpha}$ and $\hat{\beta}$. Under conditions of perfect competition and profit maximization, marginal productivity will equal the real wage. If these conditions hold, $\hat{\alpha}$ will capture labour's share of total income and $\hat{\beta}$ will capture capital's share, which is consistent with how α and β are calculated in the growth accounting framework (Intriligator 1978). In most econometric applications $\hat{\varepsilon}_i$ is afforded no special attention, other than that it satisfies classical assumptions of being normally distributed with a zero mean. But, analogously to the macro-level interpretation of ΔA, $\hat{\varepsilon}_i$ (or some portion of $\hat{\varepsilon}_i$) has been interpreted as capturing deviations from efficient behaviour among the organizations under scrutiny, with inefficiency defined as the extent to which an organization's output falls short of that predicted by the production function.

At the meso-level, the production function models the maximum output an organization could secure, given its level and mix of inputs. The production process is shown in very simple terms in Fig. 2.7.1. The organization employs inputs (labour, capital, equipment, raw materials) and converts them into some sort of output. The point at which this production process takes place (middle box) is critical for determining whether some organizations are better at converting inputs into outputs.

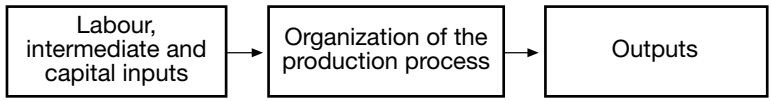

Fig. 2.7.1 Simplified production process

The middle box is something of a black box because it is usually very difficult for outsiders to observe an organization's operation and the organization of the production process. In some industries (e.g. pharmaceutical sector) the production process is a closely guarded secret and the source of competitive advantage.

This inability to observe the production process directly is a fundamental challenge for those seeking to analyse productivity or efficiency. Nevertheless, it is possible to devise a gold standard production process that describes the best possible way of organizing production, given the prevailing technology. The point at which the amount and combination of inputs is optimal is termed the production frontier – any other scale of operation or input mix would secure a lower ratio of output to input. Organizations that have adopted this gold standard are efficient, operating at the frontier of the prevailing technological process. Organizations can operate some way short of this gold standard if equipment is outmoded, the staff underperforms or capital resources stand idle periodically. These, and multiple other reasons, might explain inefficiency.

The analytical problem comprises the following challenges: the gold standard production process is unknown; the particular form of the production process adopted in each organization is difficult to observe; and the various shortcomings associated with each of these particular processes are poorly understood. These challenges can be addressed by comparing organizations involved in similar activities. Such compara-

tive analysis does not attempt to prise open the black box but concentrates on the extremes depicted in Fig. 2.7.1. Information about what goes in (inputs to production process) and what comes out (outputs of production process) tends to be available in some form or another and allows comparison of input-output combinations between organizations that produce similar things. An organization is more *productive* if it uses less input to produce one unit of output than another organization. If we want to assess organizations that produce different amounts of output, we need to make judgements about whether there are economies of scale which, in turn, relies on understanding the gold standard production process. If this is known, organizations can be judged in terms of their *efficiency*.

Distinguishing productivity and efficiency

Productivity and efficiency are often used interchangeably but they refer to different concepts. Sometimes they are distinguished according to what is measured – productivity used when output is measured by activities or services and efficiency used when output is measured by health outcomes. The OECD (2005) has separated technical (or cost) effectiveness from technical (or cost) efficiency – efficiency applies when output is measured by activities; effectiveness when output is measured by outcomes such as health gains or equity.

In country surveys the OECD distinguishes between the concepts of macro- and micro-efficiency (OECD 2003). Macro-efficiency relates to the question of whether total health expenditure is at a socially desirable level. Micro-efficiency involves either minimizing the cost needed to produce a given output or maximizing output for given costs. Within the concept of micro-efficiency, the OECD defines productivity as the volume of services per dollar of expenditure on inputs and effectiveness as quality of care, including health improvement and responsiveness (e.g. timely provision of care).

The definitions used in this chapter are given below.

- Productivity is the ratio of a measure of output to a measure of input.
- Technical efficiency is the maximum level of output that can be produced for a given amount of input under the prevailing technological process – the gold standard.

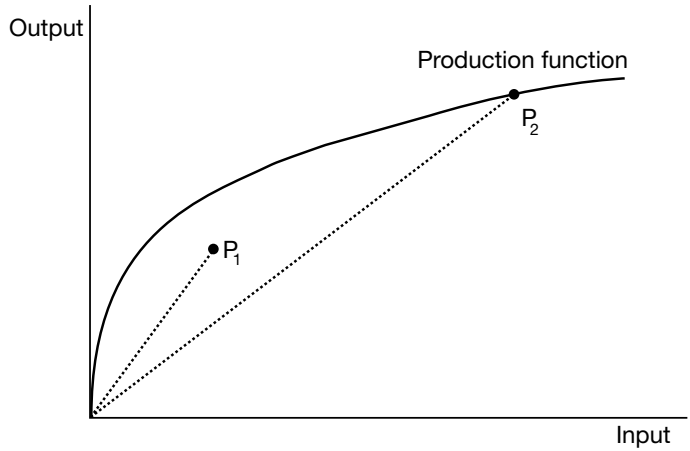

Fig. 2.7.2 Productivity and efficiency

- Allocative efficiency is the maximum level of output that can be produced assuming the cheapest mix of inputs given their relative prices.

The difference between the first two measures is shown in Fig. 2.7.2. Two organizations ($P_{1;}$ P_2) use a single input to produce a single type of output but P_1 has a higher level of productivity i.e. a higher ratio of output to input. However, technical efficiency is measured in relation to the production function – the maximum amount of output that can be produced at different levels of input. This function suggests diminishing marginal productivity – each additional unit of input produces progressively less output. Diminishing marginal productivity implies decreasing returns to scale – the more inputs used, the lower the return in the form of outputs.

In this illustration, P_2 is operating on the production function, producing the maximum level of output that is technically feasible given its input levels. In contrast, P_1 is operating inefficiently given its size – P_1 has a higher output/input ratio than P_2 but at its scale of operation it would be technically feasible to produce more output. The technical inefficiency of P_1 is measured by its vertical distance from the production function.

Organizations can be allocatively inefficient if they do not use the correct mix of inputs according to their prices. This can be illustrated

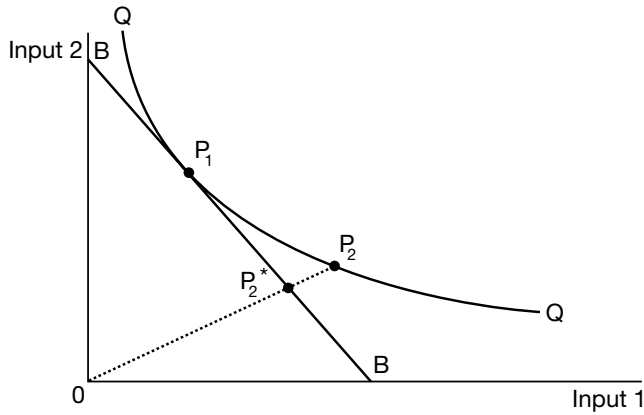

Fig. 2.7.3 Allocative efficiency with two inputs

in a simple two input model. For some known production process, the isoquant QQ in Fig. 2.7.3 shows the use of minimum combinations of the two inputs required to produce a unit of output. In this figure, the organizations P_1 and P_2 lie on the isoquant and therefore (given the chosen mix of inputs) cannot produce more outputs. They are both technically efficient. Organizations might not adopt the best combination of inputs given their prices. Suppose the market prices of the two inputs are V_1 and V_2 – the cost minimizing point on the isoquant occurs where the slope is $-V_1/V_2$ (shown by the straight line BB). In Fig. 2.7.3 this is at the point where P_1 lies, which is allocatively efficient. However, although P_2 lies on the isoquant the organization is not efficient with respect to prices, as a reduction in costs is possible. The allocative inefficiency of P_2 is given by the ratio OP_2^*/OP_2.

Organizations may exhibit both allocative and technical inefficiency. This is illustrated in Fig. 2.7.4 by comparing organizations P_3 and P_4. Organization P_3 purchases the correct mix of inputs but lies inside the isoquant QQ. It therefore exhibits a degree of technical inefficiency, as indicated by the ratio OP_1/OP_3. Organization P_4 purchases an incorrect mix of inputs (given their prices) and lies inside the isoquant QQ. Its overall level of inefficiency is measured as OP_2^*/OP_4, which comprises two components: (i) the organization's allocative inefficiency indicated by the ratio OP_2^*/OP_2; and (ii) its technical inefficiency indicated by the ratio OP_2/OP_4.

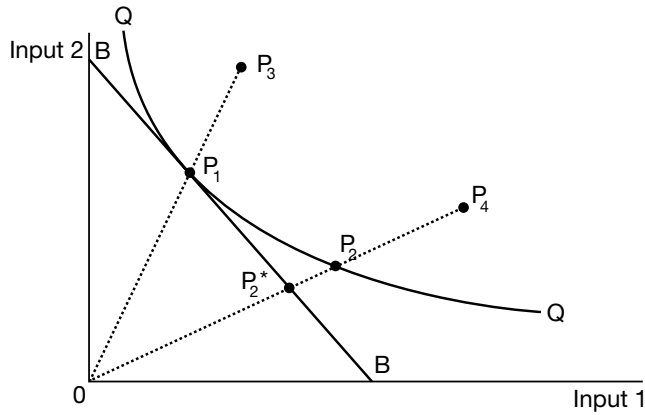

Fig. 2.7.4 Technical and allocative efficiency

Defining, measuring and valuing output

Specification of the inputs consumed and the valued outputs produced is central to the examination of any production process. Analysts usually refer to the outputs of the production process but regulators and other decision-makers are usually more interested in the outcomes produced, in terms of their impact on individual and social welfare.

Physical output is usually a traded product in competitive industries. Even in a reasonably homogeneous market, the products (e.g. cars) can vary considerably in various dimensions of quality such as reliability or safety features (Triplett 2001). The quality of the product is intrinsic to its social value but that value can be readily inferred by observing the price that people are prepared to pay. For this reason there is usually no need explicitly to consider the ultimate outcome of the product, in terms of the value it bestows on the consumer.

Prices do not exist and outputs are difficult to define in many parts of the economy. This is particularly true for many of the goods and services funded by governments (Atkinson 2006). Some of these are classic public goods (non-rival and non-excludable) that would be underprovided if left to the market, e.g. national defence. Government financing of other services (e.g. education, health care) might be justified to ensure universal access. Two fundamental issues need to be considered in the context of productivity and efficiency analysis. How should the outputs of the non-market sector be defined? What value should be attached to these outputs when market prices are not available?

Defining health outcomes

When defining health outcomes the starting point is to consider the objectives of the health system or organization(s) under consideration. The primary purpose of the health-care system is generally considered to be to enhance the health of the population. Individuals do not demand health care for its own sake but for its contribution to health. Presuming that the health system and its constituent organizations aim to satisfy individual demands (however imperfectly) it follows that health should enter the social welfare function and organizational objective functions. Ideally, the measure of health should indicate the value added to health as a result of an individual's contact with the health system. This requires a means of defining and measuring individual health profiles and of attributing changes in these to the actions of the health system or its constituent organizations.

Health is multidimensional and – like utility – there is no objective means of measuring and ordering health across individuals or populations. A diversity of definitions have been used including life expectancy; capacity to work; personal and social functioning; and need for health care (Fuchs 1987). One option is to use avoidable deaths or amenable mortality as an output measure. This is based on a list of causes of deaths that should not occur in the presence of effective and timely health care (Nolte & McKee 2003; Nolte et al. 2009). The aim is to ascertain health services' effect on mortality by disentangling other influences that are unrelated to the health system.

Data on the impact of health services on morbidity or health-related quality of life (HRQoL) are seldom collected outside of clinical trial settings and therefore have rarely been used in productivity analyses. This may change as more countries start to collect such data, even from patients who are not enrolled in clinical trials (Department of Health 2007; Räsänen 2007; Vallance-Owen et al. 2004).

Defining the quantity of output

Given the current absence of data on the amount of health produced, most productivity analyses define output in terms of the numbers and types of patients treated. Sometimes they adjust for the quality of treatment. This is in line with a common approach in theoretical expositions wherein the particular interest is often the analysis of situations in which quality substitutes for quantity (Chalkey & Malcomson 2000, Hodgkin & McGuire 1994). Consistent with such theoretical

models, Eurostat's guidance for the compilation of national accounts for European Union countries defines health-care output as: 'the quantity of health care received by patients, adjusted to allow for the qualities of services provided, for each type of health care' (Eurostat 2001).

It is difficult to define even the quantity of health care. This involves consideration of many diverse activities as the production of health care is complex and individually tailored. Contributions to the care process often come from multiple agents or organizations; a package of care may be delivered over multiple time periods and in different settings; and the responsibilities for delivery may vary from place to place and over time. This means that the production of the majority of health-care outputs rarely conforms to a production-line type technology in which clearly identifiable inputs are used to produce a standard type of output (Harris 1977).

Patient classification systems have been developed to address this problem. Patients are described reasonably well in the hospital sector as many countries use some form of diagnosis related groups (DRGs) to quantify hospital activity and to describe the different types (casemix) of patient receiving inpatient care (Fetter et al. 1980). DRGs are best suited to describe patients in hospital settings, where patients tend to be admitted with specific problems that can be managed as discrete events. Casemix adjustment methods for patients treated in outpatient, primary or community care settings are still at the development stage, although a number of classification systems are being explored (Bjorkgren et al. 1999; Carpenter et al. 1995; Duckett & Jackson 1993; Eagar et al. 2003; Street et al. 2007). A major challenge is that many patients treated in these settings have complex health-care requirements and may suffer from multiple problems that require ongoing contact with multiple agencies over a long period. Patients can be tracked across settings in countries that use unique personal identification numbers (Linna & Häkkinen 2008). Elsewhere, activity is described in fairly crude terms, such as number of attendances; or visits or consultations by setting or professional group.

Defining the quality of output

Quantity is difficult to define but it is even more challenging to assess the quality of health care. The majority of empirical studies of the efficiency of health-care organizations fail to consider quality and include only measures of casemix-adjusted quantity (Hollingsworth et

al. 1999). In effect, this assumes that there are no differences or variations over time in the quality of treatment among the organizations under consideration.

However, quality improvements are likely to be of value to patients and therefore an important aspect of health-care productivity. As mentioned, health care's impact on health status is of primary interest. Various productivity analyses have attempted to quantify improvements over time in both the amount and quality of treatment, often by considering specific conditions. For example, Shapiro and Shapiro (2001) argue that the value of cataract extraction has risen steadily because of lower rates of complication and better post-operative visual outcomes; Cutler et al. (2001) consider improvements in survival rates following treatment for heart attack; and Castelli et al. (2007) show how improvements in post-operative survival can be incorporated into measures of productivity for the whole health system.

Patients are concerned not only with the outcomes associated with care but also about the process of health-care delivery, such as the reassurance and guidance they receive; waiting times for treatment; and whether they are treated with dignity and respect. It is likely that the process of care delivery also has improved in most countries over time. These improvements ought to be included in measures of health service productivity, insofar as they represent valued improvements in the characteristics of health-care activity. This requires each dimension of quality to be measured consistently over time and a means of valuing unit changes in quality and in quantity on the same valuation scales to enable quality change to be incorporated directly in the output index. It is challenging to value both the quantity and quality of health care.

Valuing outputs

Hospital treatment following cardiac arrest has a different value to a general practitioner consultation about back pain. But how are these values to be derived in the absence of market prices? One source of valuation is based on what these activities contribute to patient welfare. This might be estimated by undertaking discrete choice experiments (Ryan et al. 2004) or by using hedonic methods to assess the value of different characteristics of outputs (Cockburn & Amis 2001). In practice, these approaches are costly and difficult to apply comprehensively across all health-care activities or to update on a routine basis.

Eurostat recommends using cost to reflect the value of non-market outputs in the national accounts (Eurostat 2001). This implies that costs reflect the marginal value that society places on these activities and requires health-care resources to be allocated in line with societal preferences (i.e. health system is allocatively efficient). This strong assumption may not hold but cost-weights have the advantage of being reasonably easy to obtain. As such, costs are likely to remain the dominant source of explicit value weights for the foreseeable future, implying that outputs are valued in terms of their production rather than consumption characteristics.

Defining inputs

The input side of efficiency analysis is usually considered to be less problematic but two issues must be faced. First, how precisely can inputs be attributed to the production of particular outputs? Second, how precisely do specific types of input need to be specified?

Attribution to the unit of analysis (i.e. the organization under consideration) is a serious analytical problem. Rather than taking the organizational form (e.g. hospital) as given, greater insight might be gained from analysing units within it, such as departments or specialties. Comparative analysis at department level makes it more likely that similar production processes are compared and may result in more robust conclusions about relative performance (Olsen & Street 2008).

Disaggregated analysis raises the question of whether it is possible to identify precisely which inputs produce which outputs. This is particularly true in health care as output is often the product of teamwork – sometimes involving collaboration between different organizational entities – and inputs (notably staff) often contribute to the production of different types of output. For instance, one doctor's time may be split between caring for patients in general surgery and in urology; another may work predominantly in dermatology but have a special interest in plastic surgery. Even the managers of the relevant specialties may not know precisely how these doctors divide their time. Ultimately, the analyst has to make a trade-off: specifying the production unit as precisely as possible (disaggregation), may come at the cost of incorrect attribution of inputs to the production process of interest.

As regards the second issue, physical inputs can be measured more accurately than outputs, or are summarized into a single measure in

the form of a measure of costs. If costs are used, a cost function can be estimated instead of a production function. The cost function indicates the minimum that an organization can incur in seeking to produce a set of valued outputs. The production function will be equivalent to the cost function (i.e. its dual) if organizations are cost minimizing – which may not be valid if the analytical purpose is to uncover inefficient behaviour. The cost function combines all inputs into a single metric (costs) but does not model the mix of inputs employed or their prices. Therefore, notwithstanding its practical usefulness, a cost function offers little help with detailed understanding of the input side of efficiency.

If there is interest in considering the impact of particular types of input on productivity, these inputs must be specified separately. In particular, separation of labour and capital may be necessary to determine their specific contributions to output (Inklaar et al. 2005).

Labour inputs

Labour inputs usually can be measured with some degree of accuracy. Most health systems collect staffing data, usually by staff type and sometimes by grade, skill level or qualifications. Care must be taken to ensure that such data are strictly comparable as organizations that report different staffing levels may actually have similar inputs. A common reason for this is varying amounts of contracting out of non-clinical (e.g. catering, cleaning, laundry services) and clinical services (laboratory, radiology). Organizations that contract out report lower staffing levels than those that employ staff directly. Differences in employment practices may also affect international comparisons. For instance, in countries such as the United States and Canada doctors are not reimbursed via the hospital and so their input may not be included in the hospital's labour statistics.

More precisely specified data may be useful if there is interest in the relationship between efficiency and the mix of labour inputs employed. This might yield useful policy recommendations about substituting some types of labour for others. But, unless there is a specific interest in the deployment of different labour types, it may be appropriate to construct a single measure of labour input – weighting the various labour inputs by their relative wages. This leads to a more parsimonious model.

Labour inputs may be measured in either physical units (hours of labour) or costs of labour, depending on context. The use of physical inputs

fails to capture any variations in organizations' wage rates. This may be desirable (e.g. if there are variations in pay levels beyond the control of organizations) or undesirable (if there is believed to be input price inefficiency in the form of different pay levels for identical workers).

Capital inputs

It is more challenging to incorporate measures of capital into the analysis. This is partly because of the difficulty of measuring capital stock and partly because of problems in attributing its use to any particular period. Measures of capital are often rudimentary and may be misleading. For example, accounting measures for the depreciation of physical stock usually offer little meaningful indication of capital consumed. Many studies of hospital efficiency use beds as a proxy for capital but this is an increasingly poor measure as care moves from inpatient to day case or other settings.

In principle, analysis should use the capital consumed in the current period as an input to the production process but, by definition, capital is deployed across time. Contemporary output may rely on capital investment in previous periods while some current activities are investments that are intended to contribute to future rather than contemporary outputs. Estimates of organizational efficiency will be biased if organizations differ in their (dis)investment strategies and capital use is attributed inaccurately to particular periods.

Macro-level analysis of productivity

Health system level

The key challenge in macro-level applications is to estimate changes in productivity over time. This requires the outputs produced from one period to the next to be measured and valued. In Laspeyres form, where outputs are valued in the base period (*t-1*), the change in output is measured as:

$$\Delta Y = Y_t - Y_{t-1}$$
$$= (outputs_t \times value\_per\_output_{t-1}) - (outputs_{t-1} \times value\_per\_output_{t-1})$$

Changes in inputs can be measured in a similar fashion. If output growth exceeds input growth it is interpreted as an improvement in productivity. However, cross-country comparisons of productivity

based on national accounts should be made with caution. Some countries (notably the United States and Canada) continue to apply the output=input convention in which the output of the health system is valued simply by the total expenditure on inputs. This makes it impossible to measure productivity because output is not measured.

Many countries have accepted Eurostat's recommendations to move towards direct measurement of the volume of outputs when constructing their national accounts (Eurostat 2001). However, there are differences in how outputs are defined in those countries that have adopted this recommendation. Many countries define health-care output by counting the number of activities undertaken in different settings – for instance, the number of patients treated in hospital or the number of attendances in outpatient departments. There is no international standard for the way that patients are described and sometimes output definitions are more akin to input measures – such as the use of occupied bed days to count the output of nursing homes or rehabilitation services. Such definitional differences undermine international comparisons (Smith & Street 2007).

A recent study developed a weighted output index to measure changes in the volume of services weighted by health gains (in quality-adjusted life years – QALYs) (Castelli et al. 2007a). No data are currently available to enable a comprehensive index to be calculated for the whole health system but the study indicates where future routine data collection should be focused.

3.2 Disease oriented approach

A number of authors have championed disease-specific assessments of productivity, often undertaken at national level (Cutler et al. 2001). They offer several potential advantages. A more focused assessment has less diversity in the type of activities being considered which simplifies their quantification and aggregation into a single index. A disease-based approach is also more likely to consider health effects and is more clearly a bottom-up approach in which micro-level comparative data on clinical actions, costs and outcomes are essential elements. They may also enable identification of specific aspects of quality change and health gain that can be overlooked when constructing a comprehensive index.

As when considering departments within organizations, there is a particular problem with identifying and attributing the resources devoted to treatment of a particular disease. This disease-based approach also presumes that it is possible to consider each disease in isolation although this may be questionable for conditions associated with multiple co-morbidities (Terris & Aron 2009). Of course, disease-specific productivity assessments should not be extrapolated to draw inferences about the productivity of the health system as a whole.

The disease-oriented approach is based on modelling the natural progress of a disease, with specific interest in the health services' role as a determinant of this progress. The idea is that analyses of time trends and more detailed (particularly individual level) data pertaining to specific health conditions will illuminate the interconnected aspects (i.e. financing, organizational structures, medical technology choices) responsible for health system performance (i.e. health outcomes and expenditure).

Most analyses are undertaken at a national level but there have been three international attempts to apply the disease-based approach during recent years.

1. McKinsey health-care productivity study – breast cancer, lung cancer, gallstone disease, diabetes mellitus: Germany, United Kingdom, United States (McKinsey Global Institute & McKinsey Health Care Practice 1996).
2. OECD Ageing-Related Disease (ARD) Project – ischaemic heart disease, stroke, breast cancer (OECD, 2003a).
3. Technological Change in Healthcare (TECH) Global Research Network (AMI) (McClellan et al. 2001).

The three projects had different perspectives. The McKinsey study analysed productivity, relating outputs (life years saved and estimations of changes in QALYs using information on mortality, complications and treatment patterns) to the resource inputs (physician hours, nursing hours, medication, capital, etc) for treating the four diseases. The study used data available at aggregate national level derived from literature reviews, database analysis and clinical expert interview. The data were limited in key areas such as clinical characteristics and detailed input measurement.

The OECD ARD Project extended the approach by trying to take account of all relevant interrelationships in a broad model. The aim was

to provide a holistic innovative framework to understand performance rather than a comparison of the countries' relative productivity. Cost and outcome data were collected on prevention, treatment and rehabilitation; the overall burden of disease; economic incentives; economic conditions; and medical knowledge. The project was implemented by collaborative networks of the participating national experts and represents the first full-scale attempt to use national micro-datasets on national patient records to compute comparable cross-sectional data. In this respect, the project can be seen as a feasibility study to examine what relevant information was available in different countries (Moise 2001). However, patient-level data on well-defined and casemix-adjusted episodes were not available so consideration of outcomes was rudimentary.

The TECH Network's aim was to study the variation in medical technology diffusion; the policy determinants of differing patterns; and the resulting consequences for health outcomes in developed countries. The Network consists of clinicians, health economists and policy-makers from seventeen nations. They have developed a multinational, standardized summary data set of acute myocardial infarction patients to analyse heart attack procedure utilization; the patient co-morbidity burden; mortality; and demographic characteristics over time and across nations. The data limitations were formidable as most of the participating countries could produce only unlinked event-based administrative or observational data. Longitudinally linked person-based data could be obtained from only seven countries.

Many challenges must still be overcome before reliable comparative studies can be undertaken across countries. Firstly, each disease will require an internationally comparable clinical protocol for measuring an episode to be defined. This should set out inclusion criteria (for example, first-ever cases); definitions of the beginning and end (follow-up) of an episode; and definitions of outcome measures. Secondly, comparable information for measuring inputs and cost must be collected, likely in several stages (Mogyorosy & Smith 2005): identification of resource items used to deliver particular services; selection of the unit of measurement of each resource item; measurement of resource items in natural units; ascribing monetary value to resource items; and expressing results in a single currency.

The disease-based approach is attractive for international productivity analysis but its usefulness is dependent on the following.

- Possibility of linking hospital discharge register to other databases. This requires a unique personal identification number and the legal possibility (confidentiality constrictors) to perform linkages.
- Availability of comprehensive register data. Register-based data are usually available for inpatient care but not primary care and the use of drugs. Hence the data are most useful for well-defined acute conditions (e.g. acute myocardial infarction, stroke) but not chronic conditions (e.g. diabetes).
- Possibility of obtaining good quality comparative input and cost data. In the ARD project, reservations have been expressed about the quality of cost data (Triplett 2002) collected from available administrative data on expenditure, costs and charges (Moise & Jacobzone 2003). The vignette method developed for international comparison of inpatient care is too crude for a disease-based approach since it is based on costing some typical cases. A better option will be to explore the methods developed for gathering comparable cost data for economic evaluations conducted on a multinational basis (Wordsworth et al. 2005) in order to meet the many challenges related to costing (Mogyorosy & Smith 2005).

Meso-level analysis of organizational efficiency

Productivity and efficiency analysis is generally conducted at organizational level. Health-care organizations use costly inputs (labour, capital, etc.) to produce valued outputs. Analysis is concerned with measuring the competence of this production process and relies on comparison of organizations that produce a similar set of outputs. If inefficiency can be revealed, it may be possible to improve the provision of health services without the need for additional resources. A number of challenges are associated with measuring organizational efficiency. The following are discussed in more detail below:

- defining comparable organizations
- identifying the production frontier
- controlling for exogenous production constraints.

Defining comparable organizations

Relative efficiency analysis requires comparison of organizations engaged in similar production processes. This is especially difficult

in contexts where the production process is characterized by varying degrees of vertical integration. It is particularly important to ensure that the entire production process is being analysed when several organizations are involved. Variations in the boundaries that define relative contributions to joint production may be a major reason why organizations have differing efficiency. For example, consider an analysis of the efficiency of care delivered to patients with head injury. The organization of care between the trauma and orthopaedics (T&O) department and the intensive care unit (ITU) may differ substantially between hospitals – some T&O departments have more step-down high dependency beds in order to relieve pressure on the ITU. If the unit of analysis is confined to the T&O department and the ITU's contribution is ignored, T&O departments that have made greater investments in high dependency beds will appear relatively inefficient although in reality they will have a better joint production process. This illustrates why sound inferences about relative efficiency cannot be made unless the analyst compares like with like.

Identifying the production frontier

As mentioned earlier, the gold standard or technically feasible production frontier is unknown. Analysis relies on estimation of an empirical frontier based on observed behaviour. Two main analytical techniques are available to assess efficiency – data envelopment analysis (DEA) and stochastic frontier analysis (SFA) (Jacobs et al. 2006).

DEA and SFA use different approaches to establish the location and shape of the production frontier and to determine each organization's location in relation to the frontier. SFA takes an indirect approach by controlling for supposed influences on output and contending that unexplained variations in output are due to inefficiency, at least in part. Standard econometric models are concerned with the explanatory variables but SFA models extract organization-specific estimates of inefficiency from the unexplained part of the model – $\hat{\varepsilon}_i$ (see Box 2.7.1). The implication is that standard econometric tools to test model specification cannot be applied to SFA models because of the interpretation placed on $\hat{\varepsilon}_i$ and because organization-specific rather than average estimates are required. This requires untestable judgments to be made about the adequacy of stochastic frontier models and the inefficiency estimates they yield (Smith & Street 2005).

DEA establishes the location and shape of the frontier empirically. The outermost observations (those with the highest level of output given their scale of operation) are deemed efficient. In Fig. 2.7.2, both P_1 and P_2 would be considered fully efficient under DEA; under SFA both organizations might be considered to exhibit some degree of inefficiency. DEA is highly flexible –by plotting the outermost observations the frontier moulds itself to the data. However, this has the drawback of making the frontier sensitive to organizations that have unusual types, levels or combinations of inputs or outputs. These will have a scarcity of adjacent reference observations and may result in sections of the frontier being positioned inappropriately.

The flexibility of DEA might be thought to increase its value over the SFA method but this is offset by two key differences in how these techniques interpret any distance from the frontier. Firstly, DEA assumes correct model specification and that all data are observed without error; SFA allows for the possibility of modelling and measurement error. Consequently, even if the two techniques yield an identical frontier, the SFA efficiency estimates are likely to be higher than those produced by DEA. Secondly, DEA uses a selective amount of data to estimate each organization's efficiency score. It generates an efficiency score for each organization by comparing it only to peers that produce a comparable mix of outputs. This has two implications.

1. Any output that is unique to an organization will have no peers with which to make a comparison, irrespective of the fact that it may produce other common outputs. An absence of peers results in the automatic assignation of full efficiency to the organization under consideration.
2. When assigning an efficiency score to an organization that does not lie on the frontier, only its peers are considered. Information pertaining to the remainder of the sample is discarded.

In contrast, SFA appeals to the full sample information to estimate relative efficiency and (in addition to making greater use of the available data) makes the sample's efficiency estimates more robust in the presence of outlier observations and atypical input/output combinations. But this advantage over DEA is mainly a matter of degree – the location of (sections of) the DEA frontier may be determined by outliers, but outliers also exert influence on the position of the SFA frontier. Moreover, there are no statistical criteria for sorting these

unusual observations into outliers or examples of best practice (Smith & Street 2005).

Controlling for exogenous production constraints

In Chapter 3.3 Terris and Aron (2009) emphasize that many factors might influence the observed performance of an organization and the importance of these situational factors is often under-emphasized. These factors may influence the organization's production frontier and constrain the amount of output it is able to produce for a given level of input. The frontiers for organizations operating in difficult situations will lie inside those of more favourably endowed organizations. For instance, hospital performance may be related to local socio-economic conditions or the organization of community care.

There is considerable debate about which situational factors are considered to be controllable. An analyst's choice will depend on whether the purpose of the analysis is short run and tactical or longer run and strategic. In the short run, many factors are outside the control of an organization; in the longer term a broader set of factors is potentially under an organization's control but the extent and nature of this control will vary with the context. In whatever way the uncontrollable environment is defined, it is usually the case that some organizations operate in more adverse situations than others, that is – external circumstances make it more difficult to achieve a given level of attainment.

Opportunities for meso-level efficiency analysis

The main requirements for meso-level analysis are that the organizations are comparable and outputs are defined in such way that the patient casemix can be standardized. At present, hospitals (or their departments) and nursing homes are most commonly studied as they meet these requirements most closely (Häkkinen & Jourmard 2007). Moreover, information systems are usually most sophisticated in the hospital sector and hospital level discharge data are available in many countries. Unique personal identification numbers allow patients to be followed along their care pathways and enable quality measures (e.g. readmission, complication, mortality) to be included in analyses (Carey & Burgess 1999; McKay & Deily 2005 & 2007).

Conclusion

Productivity and efficiency analyses consider the use of health-care resources and whether there is scope for better utilization. Productivity and efficiency have been defined in this chapter, noting that the former is a measure of the ratio of output to input while the latter incorporates the concept of what level of production might be technically feasible.

There are major challenges in measuring productivity and efficiency in health care, whether measuring the whole health system; organizations within it; or specific types of disease. The most significant challenges relate to the measurement of output although there has been much development, including improved categorization of patients and increased availability of register-based data which enable patients to be tracked over time and across settings. However, there is still a lack of routine data about health-care's impact on health outcomes and the moves to address this deficiency are to be encouraged.

Productivity analysis at health system level is often undertaken to inform national accounts and has been designed for a variety of analytical and policy purposes (macro-economic management; assessing overall economic performance and welfare). One explicit aim has been to develop measures of productivity in the health sector and its subsectors that can be compared with other sectors in the economy. The adoption of direct volume measurement has improved what is captured in the national accounts (OECD 2001). Nevertheless, there is some way to go before these accounting measures fully capture changes in health system productivity over time and enable sound international comparisons. Methodological challenges include the measurement of health outcomes, how to quantify and value outputs and how to account for quality change (Smith & Street 2007).

A disease-based approach may provide useful insight, especially if it allows analysis of health gain. Moreover, the development of electronic patient record systems may make it feasible to construct care pathways for patients who receive care from multiple providers over extended time periods. For comparative purposes, standardized definitions of activities and classifications describing the treatments (i.e. diagnosis, procedures) are required. There are analytical challenges concerning attribution, notably how to deal with co-morbidities and how to identify the resources devoted to a specific disease.

Numerous studies have considered the efficiency of health-care organizations, employing empirical techniques to make comparative statements about relative performance. Studies have become more sophisticated over time as better data have allowed improved specification of the production process; greater consideration of the quality of output; and better understanding of the situational factors that may act as constraints on production. Despite these improvements these analyses have limited impact on policy and practice, mainly because of concerns about reliability (Hollingsworth & Street 2006). Greater confidence can be gained by undertaking sensitivity analysis; estimating confidence intervals; and, most importantly, by cautious interpretation of results.

Given the fundamental analytical challenges described in this chapter, rather than claiming that inefficient behaviour can be identified precisely, we should be pursuing the more modest ambition of sorting the inefficient from the efficient. Migration from the first group to the second can then be encouraged by applying regulatory pressure; designing financial incentives; or simply sharing examples of best practice. By systematically detailing the use of resources, productivity and efficiency analyses can contribute to better targeted policy-making.

References

Atkinson, T (2006). 'Measurement of government output and productivity.' *Journal of the Royal Statistical Society, Series A,* 169(4): 659–662.

Bjorkgren, MA. Hakkinen, U. Finne-Soveri, UH. Fries, BE (1999). 'Validity and reliability of Resource Utilization Groups (RUG-III) in Finnish long-term care facilities.' *Scandinavian Journal of Public Health,* 27(3): 228–234.

Carey, K. Burgess, JF Jr. (1999). 'On measuring the hospital cost/quality trade-off.' *Health Economics,* 8(6): 509–520.

Carpenter, GI. Main, A. Turner, GF (1995). 'Casemix for the elderly inpatient: resource utilization groups (RUGs) validation project.' *Age and Ageing,* 24(1): 5–13.

Castelli, A. Dawson, D. Gravelle, H. Jacobs, R. Kind, P. Loveridge, P. Martin, S. O'Mahony, M. Stevens, P. Stokes, L. Street, A. Weale, M (2007). 'A new approach to measuring health system output and productivity.' *National Institute Economic Review,* 200(1): 105–117.

Castelli, A. Dawson, D. Gravelle, H. Street, A (2007a). 'Improving the measurement of health system output growth.' *Health Economics,* 16(10): 1091–1107.

Chalkey, M. Malcomson, J (2000). Government purchasing of health services. In: Culyer, AJ. Newhouse, JP (eds.) *Handbook of health economics*. North Holland: Elsevier.

Cockburn, IM. Amis, AH (2001). Hedonic analysis of arthritis drugs. In: Cutler, DM. Berndt, ER (eds.). *Medical care output and productivity.* Chicago: University of Chicago Press.

Cutler, DM. McClellan, M. Newhouse, JP. Remler, D (2001). Pricing heart attack treatments. In: Cutler, DM. Berndt, ER (eds.). *Medical care output and productivity.* Chicago: University of Chicago Press.

Department of Health (2007). *Guidance on the routine collection of patient reported outcome measures (PROMs)*. London: Department of Health.

Duckett, S. Jackson, T (1993). 'Casemix classification for outpatient services based on episodes of care.' *Medical Journal of Australia,* 159(3): 213–214.

Eagar, K. Gaines, P. Burgess, P. Green, J. Bower, A. Buckingham, B. Mellsop, G (2003). 'Developing a New Zealand casemix classification for mental health services.' *World Psychiatry,* 3(3): 172–177.

Eurostat (2001). *Handbook on price and volume measures in national accounts*. Luxembourg: Office for Official Publications of the European Communities.

Fetter, RB. Shin, YB. Freeman, JL. Averill, RF. Thompson, JD (1980). 'Case mix definition by diagnosis-related groups.' *Medical Care,* 18 (Suppl. 2): 1–53.

Fuchs, VR (1987). Health economics. In: Eatwell, J. Milgate, M. Newman, P (eds.). *The new Palgrave: a dictionary of economics.* London: Macmillan Press Limited.

Harris, JE (1977). 'The internal organisation of hospitals: some economic implications.' *Bell Journal of Economics*, 8(2):467–482.

Hodgkin, D. McGuire, TG (1994). 'Payment levels and hospital response to prospective payment.' *Journal of Health Economics,* 13(1): 1–29.

Hollingsworth, B. Street, A (2006). 'The market for efficiency analysis of health care organisations.' *Health Economics,* 15(10): 1055–1059.

Hollingsworth, B. Dawson, PJ. Maniadakis, N (1999). 'Efficiency measurement of health care: a review of non-parametric methods and applications.' *Health Care Management Science,* 2(3): 161–172.

Inklaar, R. O'Mahony, M. Timmer, M (2005). 'ICT and Europe's productivity performance – industry-level growth account comparisons with the United States.' *Review of Income and Wealth,* 51(4): 505–536.

Intriligator, MD (1978). *Econometric models, techniques and applications.* Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Jacobs, R. Smith, PC. Street, A (2006). *Measuring efficiency in health care: analytical techniques and health policy.* Cambridge: Cambridge University Press.

Jorgenson, DW. Griliches, Z (1967). 'The explanation of productivity change.' *Review of Economic Studies,* 34(3): 249–283.

Linna, M. Häkkinen, U (2008). Benchmarking Finnish hospitals. In: Blank, J. Valdmanis, V (eds.). *Evaluating hospital policy and performance: contributions from hospital policy and productivity research*. Oxford: Elsevier.

McClellan, M. Kessler, D. Saynina, O. Moreland, A. TECH Research Network (2001). 'Technological change around the world: evidence from heart attack care.' *Health Affairs (Millwood),* 20(3): 25–42.

McKay, NL. Deily, ME (2005). 'Comparing high- and low-performing hospitals using risk-adjusted excess mortality and cost inefficiency.' *Health Care Management Review,* 30(4): 347–360.

McKay, NL. Deily, ME (2007). 'Cost inefficiency and hospital health outcomes.' *Health Economics,* 17(7):833–848.

McKinsey Global Institute & the McKinsey Health Care Practice (1996). *Health care productivity.* Los Angeles: McKinsey and Co., Inc.

Mogyorosy, Z. Smith, PC (2005). *The main methodological issues in costing health care services – a literature review,* York, University of York: Centre for Health Economics.

Moise, P (2001). *Using hospital administrative databases for a disease-based approach to studying health care systems*. Paris: OECD.

Moise, P. Jacobzone, S (2003). *OECD study of cross-national differences in the treatment, costs and outcomes of ischaemic heart disease*. Paris: OECD.

Nolte, E. McKee, M (2003). 'Measuring the health of nations: analysis of mortality amenable to medical care.' *British Medical Journal,* 327(7424): 1129–1132.

Nolte, E. Bain, CM. McKee, M (2009). Population health. In: Smith, PC. Mossialos, E. Papanicolas, I. Leatherman, S (eds.). *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press.

Olsen, KR. Street, A (2008). 'The analysis of efficiency among a small number of organisations: how inferences can be improved by exploiting patient-level data.' *Health Economics,* 17(6): 671–681.

OECD (2001). *OECD productivity manual: a guide to the measurement of industry-level and aggregate productivity growth*. Paris, Organisation for Economic Co-operation and Development.

OECD (2003) *Ad hoc group on the OECD health project. Assessing the performance of health-care systems: a framework for OECD*

surveys. Unpublished report for official OECD use (ECO/CPE/WP1[2003]10).

OECD (2003a). *A disease-based comparison of health systems: what is best and at what cost?* Paris: OECD.

Räsänen, P (2007). *Routine measurement of health-related quality of life in assessing cost-effectiveness in secondary health care*. Helsinki: STAKES (Research Report no. 163).

Ryan, M. Odejar, M. Napper, M (2004). *The value of reducing waiting time in the provision of health care: a review of the evidence*. Aberdeen: Health Economics Research Unit.

Shapiro, I. Shapiro, MD (2001). Measuring the value of cataract surgery. In: Cutler, DM. Berndt, ER (eds.). *Medical care output and productivity*. Chicago: University of Chicago Press.

Smith, PC. Street, A (2005). 'Measuring the efficiency of public services: the limits of analysis'. *Journal of the Royal Statistical Society Series A*, 168(2): 401–417.

Smith, PC. Street, A (2007). 'Measurement of non-market output in education and health.' *Economic and Labour Market Review,* 1(6): 46–52.

Street, A. Vitikainen, K. Bjorvatn, A. Hvenegaard, A (2007). *International literature review and information gathering on financial tariffs*. York: University of York, Centre for Health Economics (Research Paper 30).

Terris, DD. Aron, DC (2009). Attribution and causality in health-care performance measurement. In: Smith, PC. Mossialos, E. Papanicolos, I. Leatherman, S (eds.) *Performance measurement for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press.

Triplett, JE (2001). What's different about health? Human repair and car repair in national accounts and national health accounts. In: Cutler, DM. Berndt, ER (eds.). *Medical care output and productivity*. Chicago: University of Chicago Press.

Triplett, JE (2002). *Integrating cost-of-disease studies into purchasing power parities (PPP)*. Washington: The Brookings Institution.

Vallance-Owen, A. Cubbin, S. Warren, V. Matthews, B (2004). 'Outcome monitoring to facilitate clinical governance: experience from a national programme in the independent sector.' *Journal of Public Health,* 26(2): 187–192.

Wordsworth, S. Ludbrook, A. Caskey, F. Macleod, A (2005). 'Collecting unit cost data in multicentre studies. Creating comparable methods.' *European Journal of Health Economics,* 6(1): 38–44.